UNIVERSIDADE FEDERAL DE SANTA CATARINA
CURSO DE PÓS-GRADUÇÃO EM
ESTUDOS DA TRADUÇÃO

Paula de Paiva Villasbôas

Análise das correspondências de tradução inglês-português para substantivos e adjetivos compostos hifenizados da língua inglesa: uma abordagem de base em *corpus*

FLORIANÓPOLIS

2009

UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO DE COMUNICAÇÃO E EXPRESSÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM
ESTUDOS DA TRADUÇÃO

Paula de Paiva Villasbôas

Análise das correspondências de tradução inglês-português para substantivos e adjetivos compostos hifenizados da língua inglesa: uma abordagem de base em *corpus*

Dissertação apresentada ao curso de Pós-Graduação em Estudos da Tradução da Universidade Federal de Santa Catarina como parte dos requisitos para obtenção do grau de Mestre em Estudos da Tradução.

Orientador: Dr. Marco Rocha

Florianópolis
2009

Catalogação na fonte pela Biblioteca Universitária da
Universidade Federal de Santa Catarina

V726 Villasboas, Paula de Paiva
Análise das correspondências de tradução inglês-português para substantivos e adjetivos compostos hifenizados da língua inglesa [dissertação] : uma abordagem de base em corpus / Paula de Paiva Villasbôas ; orientador, Marco Rocha. – Florianópolis, SC : 2009.
80 f.: il., tabs.

Dissertação (mestrado) – Universidade Federal de Santa Catarina, Centro de Comunicação Expressão. Programa de Pós-Graduação em Estudos da Tradução.

Inclui bibliografia

1. Linguística. 2. Língua inglesa – Tradução mecânica. 3. Língua portuguesa – Tradução mecânica. I. Rocha, Marco. II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Estudos da Tradução. III. Título.

CDU 801=03

Dissertação julgada para a obtenção de grau de
MESTRE EM ESTUDOS DA TRADUÇÃO
Área de concentração: Processos de Retextualização

Lexicografia, tradução e ensino de línguas.

Aprovada em sua forma final pelo
Programa de Pós-Graduação em Estudos da Tradução
da Universidade Federal de Santa Catarina

BANCA EXAMINADORA:

Prof. Dr. Marco Antônio Esteves da Rocha
Orientador

Profa. Dra. Stella Tagnin
(USP)

Prof. Dr. Markus J. Weininger
(UFSC)

Prof. Dr. Walter Carlos Costa
Suplente

Profa. Dra. Andréia Guerini
Coordenadora

AGRADECIMENTOS

Várias pessoas foram imprescindíveis à realização desta dissertação, entre elas:

- Luís Romero Jardim Villasbôas, meu pai, e Clarisse Villasbôas minha mãe, pela compreensão e apoio financeiro;
- Meu irmão, Rodrigo de Paiva Villasbôas, pela solicitude demonstrada principalmente na realização da análise quantitativa do trabalho;
- Minha irmã e meu cunhado, Andrea Villasbôas Malburg e Túlio Eugênio Malburg, por mais uma vez me incentivarem e demonstrarem interesse pela minha pesquisa;
- Aos meus saudosos amigos Simone Santos Guimarães e Vítor Guimarães, e sua filhinha Isadora Guimarães, que me acolheram em Florianópolis tanto na qualificação como na defesa final de minha dissertação;
- Meus saudosos amigos Neres Bitencourt e Eduardo Juan Soriano-Serra, pelo apoio e pelos momentos agradáveis proporcionados durante minha estada em Florianópolis;
- Minhas saudosas amigas Josiane Adam e Márcia Machado, pelo incentivo e interesse demonstrado pelo trabalho;
- Adriane Bitencourt, minha amiga e diretora de um escritório de tradução para o qual eu trabalho;
- Meu orientador, Prof. Marco Rocha, PhD, pelo apoio e orientação competente;
- A banca da qualificação do projeto de dissertação, assim como a da defesa final, pelas contribuições feitas no sentido de melhorar meu trabalho;
- A UFSC, pelo apoio institucional.

RESUMO

Esta pesquisa busca analisar as correspondências de tradução inglês-português para substantivos e adjetivos compostos hifenizados da língua inglesa. Para tanto, utiliza uma abordagem baseada em *corpus (corpus-based)*. O *corpus* utilizado é composto por textos de um documento oficial. O objetivo principal desta análise é fornecer informação morfossintática para auxiliar a geração de regras de tradução para sistemas tradução automática. O *corpus* também é analisado para verificar e quantificar algumas das características de documentos oficiais apresentadas por Biber (1988). Os procedimentos metodológicos foram realizados por meio do programa *WordSmith Tools*, utilizando-se as ferramentas *Wordlist, Concord*, e *Keywords.* O *corpus* de referência utilizado foi o *British National Corpus* (*BNC*). A princípio foi gerada uma lista de palavras por meio da ferramenta *Wordlist*, para obter informações estatísticas, como o número de tipos e ocorrências do documento oficial. A Ferramenta *Keywords* foi utilizada para identificar as palavras em sobreuso no *corpus* de estudo, em relação ao *corpus* de referência (*BNC*). Tais palavras formaram a lista de palavras-chave do documento. Verificou-se que juntamente com algumas dessas palavras-chave ocorriam numerosos substantivos e adjetivos compostos hifenizados. Tais nomes compostos hifenizados foram analisados para verificar as estruturas morfossintáticas de cada composto em inglês e de sua tradução para o português. Foram analisados quarenta e um exemplos de tradução de nomes compostos. A análise da tradução desses exemplos produziu informação morfossintática a ser utilizada em formalismos de regras de tradução, para aplicação na tradução automática. A análise do *corpus* confirmou a maior parte das características de documentos oficiais apresentadas por Biber, e ainda revelou outras características do *corpus* de estudo, como a ocorrência de numerosos verbos modais, relacionados ao caráter prescritivo do documento. Por sua vez, a análise das correspondências de tradução para nomes compostos hifenizados forneceu informação que pode ser aproveitada em sistemas de tradução automática direta, especialmente em *corpora* de domínio restrito; ou em sistemas híbridos, de modo que cada método seja responsável por determinados aspectos do processo de tradução.

Palavras-Chave: Linguística de corpus, nomes compostos hifenizados, correspondências de tradução, tradução automática.

ABSTRACT

This research aims to analyze English-Portuguese translation correspondences for hyphenated compound nouns. A corpus-based approach was used considering texts of an official document. The main objective is to provide morphosyntatic information to aid the generation of translation rules for translation machine systems. The research also verified and quantified some characteristics of this official document, according to those presented in Biber (1988). Methodological procedures were accomplished through the software WordSmith, using *Wordlist, Concord*, and *Keywords* tools. The reference corpus was the *British National Corpus* (*BNC*). Firstly, a word list was created to get statistic information, as types and tokens numbers for this official document. The Keywords tool was used to identify the overuse of some words in the corpus, in comparison with the reference corpus (BNC). Such words composed a list of keywords for this document. Together with some of these keywords it was verified numerous hyphenated compound nouns. Such hyphenated compounds were analyzed in order to get information about the morphosyntatic structure both for the compound noun in English and for its translation into Portuguese. Forty one examples of translation for compound nouns were analyzed. This analysis provided morphosyntatic information to be used in formalisms of translation rules, for the application in machine translation. The study of the corpus confirmed a larger part of characteristics of official documents presented in Biber, and still revealed other characteristics for the corpus, as the occurrence of numerous modals, related to the prescriptive character of the document. On the other hand, the translation correspondences for hyphenated compounds contributed with information which can be used in direct machine translation systems, specially in restrict domain corpora; or in hybrid systems, so that each method can be responsible for certain aspects of the translation process.

Keywords: corpus linguistics, hyphenated compound nouns, translation correspondences, machine translation.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Embora a prática da tradução e a reflexão sobre a natureza do processo tradutório tenham iniciado há séculos atrás, nos tempo de Cícero e São Jerônimo, os Estudos da Tradução não emergiram como uma disciplina acadêmica separada até os anos 80. Foi Holmes (1988) quem propôs uma estrutura geral para o novo campo em seu artigo seminal 'The name and nature of translation studies'. Holmes e muitos outros especialistas da tradução – entre eles Lefévère e Bassnett – expressaram a mesma insatisfação com os métodos introspectivos, ou seja, baseados nas próprias idéias e intuições do tradutor, e advogaram o recurso a grandes coleções de textos traduzidos.

O aparecimento recente de uma tendência com base em *corpus* dentro dos estudos da tradução pode ser visto como consequência direta de uma principal mudança de perspectiva nesse campo, que desloca o foco de pesquisa do texto fonte para o texto alvo. Por vários anos a pesquisa focalizou a equivalência em relação ao texto fonte, até que teóricos como Toury (1980) advogaram uma orientação ao alvo e transformaram os estudos da tradução em um esforço descritivo para especificar as leis da tradução. No começo dos anos 90, Mona Baker foi quem iniciou a tendência baseada em *corpus*, coletando corpora de textos traduzidos para revelar os padrões que distinguem uma tradução da outra (GRANGER, 2003). Essa pesquisadora é a responsável pela fundamentação (com base em corpora) da existência de traços linguísticos comuns em traduções. Suas investigações trouxeram à luz o conceito de "translation universals", uma questão já levantada por Blum-Kulka (1986), e posteriormente enriquecido por Laviosa-Braithwaite (1998) e por Toury (2000). Segundo Granger (2003), estudos de *corpus* posteriores têm tentado confirmar ou refutar tais universais de tradução.

A criação e investigação sistemática de corpora têm continuamente aberto novas áreas de pesquisa em linguística descritiva e aplicada. Áreas como a linguística contrastiva, lexicografia, linguística computacional, terminologia, linguística forense, entre outras, têm experimentado mudanças significativas em suas abordagens teóricas e metodológicas, graças à influência da linguística de *corpus*.

Em particular, a união entre a pesquisa com base em corpora e os Estudos da Tradução tem gerado um paradigma híbrido, coerente e rico, que trata de uma variedade de assuntos pertencentes à teoria, descrição e prática da tradução. Esse novo paradigma, os Estudos da Tradução Baseados em *Corpora* (*Corpus-based Translation Studies*), pode ser definido como o ramo da disciplina que utiliza corpora de textos originais e/ou traduzidos para o estudo empírico do produto e processo da tradução, elaboração de construtos teóricos, e treinamento de tradutores (LAVIOSA, 2003).

Dados empíricos têm sido usados em lexicografia muito antes da disciplina de linguística de *corpus* ser inventada. Os *corpora,* no entanto, mudaram o modo pelo qual os linguistas podem examinar uma língua. A linguística de *corpus* tem se desenvolvido consideravelmente nas últimas décadas devido às grandes possibilidades oferecidas pelo uso de computadores, bem como pela disponibilidade de textos digitalizados, o que tem possibilitado obter dados rápida e facilmente e ter esses dados em um formato adequado para análise (McENERY & WILSON, 1996).

O impacto dos *corpora* na lexicografia tem sido benéfico, haja vista o grande número de editores de dicionários que estão investindo em tecnologia de *corpus.* Especialmente, os *corpora* têm tido um efeito importante e duradouro sobre o trabalho de lexicografia monolíngue. Já o impacto dos *corpora* da produção de dicionários multilingues ainda está por ser demonstrado. Isso ocorre, em parte, devido à escassez dos *corpora* paralelos em comparação ao número de *corpora* monolíngues, e também pelas dificuldades de acesso aos *corpora* multilíngues (McENERY & WILSON, 2001). Os *corpora* de textos multilíngues podem ser utilizados para esclarecer as diferenças entre os textos originais e suas traduções e para estudos de problemas e estratégias de tradução individuais (AIJMER e ALTENBERG, 2000).

Flowerdew (2004) usa os termos '*corpora* geral' e '*corpora* especializado' para diferenciar os *corpora* que consistem em uma ampla variedade de gêneros escritos e falados, daqueles que foram planejados com um propósito específico em mente: investigar o uso e estrutura da linguagem especializada. Os *corpora* especializados são mais relevantes para o propósito de entender tipos específicos da linguagem acadêmica e profissional.

Nesta pesquisa os *corpora* de estudo são compostos por textos escritos provenientes de um documento oficial da área de gestão ambiental (Agenda 21), que podem ser considerados como '*corpora* especializados' ou '*corpora* de domínio restrito'. São utilizadas as versões em

inglês e português desse documento, constituindo dessa forma um *corpus* paralelo. O *corpus* de referência para a língua inglesa é o BNC (*British National Corpus*), mas apenas os textos escritos, que são considerados como um *corpus* geral. O *corpus* de referência tem a função de fornecer uma norma com a qual será feita a comparação das frequências do *corpus* de estudo (BERBER SARDINHA, 2004).

A **hipótese** deste trabalho consiste no seguinte: "A análise da tradução de substantivos e adjetivos compostos hifenizados em inglês, inferida pelas correspondências de tradução dos mesmos para o português, pode auxiliar a geração de regras de tradução para aplicação em sistemas de tradução automática."

1.1 Objetivos

1.1.1 Objetivo Geral

Esta pesquisa procura quantificar e examinar as características de um *corpus* especializado composto por textos de um documento oficial, bem como analisar as correspondências de tradução inglês-português para substantivos e adjetivos compostos hifenizados da língua inglesa, tendo como base as diferenças na estrutura morfossintática. Desse modo, busca-se auxiliar a geração de regras de tradução para aplicação na tradução automática.

1.1.2 Objetivos Específicos

- Caracterizar o *corpus* com o uso do programa *Wordsmith Tools,* utilizando um *corpus* geral (*British National Corpus*) como *corpus* de referência para verificar as palavras-chave características do *corpus* de estudo (Agenda 21);

- Verificar a ocorrência de adjetivos e substantivos compostos hifenizados que coocorrem com algumas palavras-chave do *corpus* de estudo;
- Analisar as correspondências de tradução desses adjetivos e substantivos compostos, do ponto de vista morfossintático;
- Classificar e quantificar os tipos deste documento oficial, com a finalidade de verificar algumas das características de documentos oficiais apresentadas por Biber (1988).

2 Referencial teórico

2.1 O uso de *corpora* nos estudos da tradução

Os estudos da tradução do ponto de vista acadêmico compõem uma disciplina relativamente nova. Embora o fenômeno da tradução venha sendo estudado no meio acadêmico há muito tempo, principalmente sob a rubrica de literatura comparada ou linguística contrastiva, apenas a partir da segunda metade do século XX os estudiosos começaram a discutir a necessidade de conduzir uma pesquisa sistemática e desenvolver teorias coerentes sobre tradução (BAKER, 1998).

Os estudos da tradução com base em *corpora* representam uma área que tem atraído um número crescente de estudiosos, que acreditam genuinamente no potencial desse tipo de pesquisa para satisfazer a pluralidade de necessidades e interesses dentro dessa disciplina (LAVIOSA, 2002).

De acordo com Hunston (2002 apud BERBER SARDINHA, 2002), os *corpora* têm mais a oferecer aos tradutores do que pode parecer à primeira vista. Eles podem não somente fornecer evidências de como as palavras são usadas e quais as traduções possíveis para uma dada palavra ou frase, como também podem auxiliar no discernimento acerca do processo e da natureza da tradução em si.

Os *corpora* são coletâneas de textos escritos ou transcrições de fala muito úteis para os estudos da tradução. Quando reunidos em formato de arquivo legível por computador podem ser chamados de *corpora* eletrônicos. Berber Sardinha (2002) afirma que a utilização de *corpora* eletrônicos na tradução pode beneficiar muito tanto a prática tradutória quanto os estudos da tradução. Para Baker (1996 apud BERBER SARDINHA, 2003, p. 44) o *corpus* eletrônico é um instrumento revolucionário "que permite enxergar aspectos da linguagem do texto traduzido, em particular, de modo muito mais rico e abrangente do que por outros meios".

Para Tagnin (2002, p. 199), a escassez de recursos lexicográficos e fraseológicos é evidente, e os dicionários não conseguem acompanhar o ritmo de criação de novas palavras. Por sua vez, a busca em um *corpus* não apenas fornece unidades fraseológicas corretas, "mas

principalmente a forma mais usual na língua sob investigação". Por isso, o recurso a *corpora* é fundamental para assegurar uma tradução em linguagem usual.

Alguns autores como Laviosa (1998) afirmam que a adoção de metodologias baseadas em *corpus* tem aumentado entre estudiosos e praticantes da tradução. Bowker (2000) também cita a 'crescente popularidade' do uso de *corpora* aplicado aos estudos da tradução. No entanto, ainda podem existir alguns fatores que dificultam a utilização de *corpora* em estudos da tradução: falta de recursos (software e *corpora*), difícil acesso à infraestrutura tecnológica (computadores e redes), e problemas relacionados à capacitação pessoal (BERBER SARDINHA, 2003).

Os *corpora* paralelos são mais bem caracterizados como a "Pedra de Rosetta" da linguística de *corpus* moderna. Esse tipo de *corpora* é tipicamente bilíngue em vez de multilíngue, ou seja, uma obra traduzida para diversas línguas (McENERY E WILSON, 1993).

Verificou-se nesta pesquisa que não há consenso sobre a denominação dos *corpora*. Para Aijmer e Altenberg (2000), os *corpora* paralelos pode se constituir em: textos semelhantes em duas ou mais línguas – chamados de '*corpora* comparáveis' – *ou* textos originais e suas traduções para outra língua, chamados pelos autores de '*corpora* de tradução'. Os *corpora* paralelos têm sido reconhecidos como recursos indispensáveis para a pesquisa teórica e prática em todos os níveis de descrição linguística.

Já para Tagnin (2002), existem dois tipos de *corpora* que possuem mais utilidade. Um deles é o *corpus* paralelo, constituído de textos originais e suas respectivas traduções. O outro, chamado de *corpus* comparável, é composto de textos similares (do mesmo gênero, tipologia, extensão e data de publicação), originais, e bilíngues (nas duas línguas de trabalho do tradutor).

Esta pesquisa optou pelo termo *corpus* paralelo para denominar o tipo de *corpus* utilizado. Tal *corpus* é composto por textos originais em inglês e suas traduções para o português.

2.2 A linguística de *corpus*

Berber Sardinha (2004) descreve a linguística de *corpus* como uma área de pesquisa que lida com a coleta e exploração de *corpora*, ou conjuntos de dados linguísticos textuais coletados

segundo alguns critérios pré-definidos, para fazer pesquisa sobre uma língua ou variedade linguística.

O problema relacionado à definição da linguística de *corpus* como uma teoria ou metodologia tem sido debatido a partir de diferentes pontos de vista. Tem sido argumentado que a linguística de *corpus* não é uma área de pesquisa e sim uma base metodológica para o estudo de uma língua. Muitos linguistas que trabalham com um *corpus*, no entanto, tendem a concordar que a linguística de *corpus* vai muito além deste papel puramente metodológico. Halliday (1993 apud TOGNINI-BONELLI, 2001), por exemplo, destaca que a linguística de *corpus* reúne as atividades de coleta de dados e teorização, e argumenta que isso está levando a uma mudança qualitativa em nossa faculdade de compreender os aspectos lexicais e gramaticais relacionados à linguagem. Outros linguistas apontam para a conexão entre o uso de métodos computacionais e estatísticos e a mudança qualitativa das observações que derivam desse tipo de abordagem (TOGNINI-BONELLI, 2001).

A ferramenta mais simples e amplamente usada para a pesquisa baseada em *corpus* é o programa de concordância, que busca no *corpus* um traço linguístico específico, ou um conjunto relacionado de traços, com a finalidade de investigação. Esse instrumento é muito útil na pesquisa lexicográfica, para identificar os significados de um item e seu contexto, além de outras características (sintáticas, estilísticas, pragmáticas) relevantes ao uso do item (LEECH, 1991).

De acordo com Partington (1998) a concordância, juntamente com a intuição do pesquisador, possibilita isolar termos dos quais se suspeita que tenham um papel na dêixis textual e que fornecem o ambiente fraseológico imediato do item. O recurso de 'visualização' (*view*) dos programas concordanciadores permite ver como um determinado item lexical se ajusta dentro de um dado contexto, e encontrar o segmento de texto ao qual está sendo referido.

Segundo Berber Sardinha (2004), o fenômeno mais tradicionalmente enfocado no estudo de *corpus* é a colocação. Conforme Sinclair (1991 *apud* PARTINGTON, 1998, p. 15) "Colocação é a ocorrência, em um texto, de duas ou mais palavras dentro de um curto espaço uma da outra." [1] A coocorrência de dois itens lexicais se torna interessante quando ela parece acontecer com um propósito, e especialmente se isso é repetido, se há "padrões de colocação".

Partington apresenta ainda outras definições desse fenômeno:

---

[1] Em inglês*: Collocation is the occurrence of two or more words within a short space of each other in a text* (SINCLAIR, 1991, Tradução da Autora).

- *"O sentido colocacional consiste nas associações que uma palavra faz por conta dos sentidos das outras palavras que tendem a ocorrer no seu ambiente.[2]" (definição psicológica ou associativa)* (LEECH, 1974 *apud* PARTINGTON, 1998).

- *"Colocação tem sido o nome dado à relação que um item lexical tem com itens que aparecem com probabilidade significativa no seu contexto (textual).[3]" (definição estatística)* (HOEY 1991, *apud* PARTINGTON, 1998).

Por exemplo, o sentido da colocação '*capacity-building activities*' deriva-se da relação estabelecida entre o termo '*activities*' e o sentido do substantivo composto hifenizado '*capacity-building*' a ele associado. Nesse exemplo, o sentido de *capacity-building* especifica o nome geral '*activities*', que passa a significar uma espécie de atividade.

Este trabalho busca identificar a analisar a tradução dos substantivos e adjetivos compostos hifenizados que coocorrem com certas palavras-chave do *corpus* de estudo.

2.3 Tradução automática

Os sistemas de tradução automática (TA) podem ser classificados de acordo com seu método ou seu paradigma. Os métodos se referem ao projeto de processamento, enquanto que os paradigmas se referem aos componentes de representação do conhecimento que auxiliam o projeto de processamento global. A seguir é feita uma breve apresentação dos métodos de TA (SPECIA E RINO, 2002).

Existem dois tipos de tradução automática: direta e indireta. A TA direta transforma as sentenças da língua fonte (LF) em sentenças de língua-alvo (LA), sem utilizar representações intermediárias. Nesse método, procura-se realizar o mínimo de processamento linguístico

[2] Em inglês: *Collocative meaning consists of the associations a word acquires on account o f the meanings of words which tend to occur in its environment.* (LEECH, 1974, Tradução da Autora).

[3] Em inglês: *Collocation has long been the name given to the relationship a lexical item has with items that appear with greater than random probability in its (textual) context.* (HOEY, 1991, Tradução da Autora)

possível. Esse processamento pode incluir a simples substituição das palavras da sentença fonte por suas correspondentes na língua alvo (tradução palavra-por-palavra), bem como tarefas mais complexas, como a reordenação de palavras na sentença alvo e a inclusão de preposições. Os sistemas de TA direta são geralmente bilíngues (construídos para um único par de línguas) e unidirecionais (traduzem somente da LF para a LA) (SPECIA E RINO, 2002).

Esses sistemas apresentam problemas como a não tradução de certas palavras (por não existirem no dicionário), ou a geração de construções gramaticais desconhecidas, por não existirem regras de transformação adequadas. DORR et al. (2000 *apud* SPECIA E RINO, 2002) afirmam que se o método for aplicado para textos simples e de domínio restrito, os resultados podem ser bastante úteis, principalmente para especialistas naquele domínio, que podem fazer a pós-edição adequada do texto traduzido.

Já no que se refere a TA indireta, existem dois tipos de métodos: por transferência e por interlíngua. A TA por transferência engloba as seguintes fases: análise, transferência e geração. Na fase de análise são mais comuns sistemas que se limitam à análise sintática, que geram uma representação intermediária abstrata. Nesse caso, a fase de transferência converte essa representação abstrata da LF em uma representação da LA, por meio de regras de mapeamento entre as duas línguas naturais. Tais regras indicam as correspondências lexicais e sintáticas entre essas representações.

A TA por interlíngua tem a função de extrair a representação do significado da sentença fonte para gerar a sentença na LA. Esse método de TA consiste primeiramente na análise completa do texto na LF, extraindo seu significado e representando-o na interlíngua; para posteriormente permitir a geração do texto na LA, partindo da representação interlingual e expressando o mesmo significado. Uma das maiores dificuldades desse método é a especificação da interlíngua, que deve ser independente de qualquer língua natural para que o significado de suas sentenças seja representado uniforme e consistentemente. A interlíngua pode ser, por exemplo, baseada em uma linguagem artificial, ou um conjunto de conceitos semânticos primitivos comuns a todas as linguagens. É necessário fazer análises exaustivas sobre a semântica do domínio da tradução em questão (SPECIA E RINO, 2002; HUTCHINS, 1995).

Muitos sistemas de TA consistem em abordagens híbridas que podem incluir combinações não só entre paradigmas, mas também entre diferentes métodos, sendo que cada

método é responsável pelo tratamento de determinados aspectos da tradução (sistemas *multi-engine*).

Até meados dos anos 90, a maior parte da pesquisa em TA ainda era baseada na implementação de regras lexicais e gramaticais, o que é chamado atualmente de tradução automática baseada em regras (rule-based machine translation, RBMT). Atualmente, os paradigmas dominantes de pesquisa em TA são baseados em *corpus* (HUTCHINS, 1995).

Neste trabalho o paradigma de interesse é o da TA baseada em *corpus*, especialmente os sistemas baseados em exemplos (*Example-Based Machine Translation*, ou EBMT). Segundo Caseli (2004), a EBMT emprega o reconhecimento de um padrão para traduzir parte da sentença fonte fornecida, para desse modo, determinar a tradução. O paradigma de EBMT é muito adequado para a obtenção de regras de tradução, uma vez que os exemplos são dados reais da língua, e representam as estruturas dentro do contexto das línguas fonte e alvo.

Para a utilização de técnicas de EBMT é necessária a existência de um *corpus* paralelo alinhado – um conjunto de exemplos (geralmente sentenças, mas nesta pesquisa são utilizados como exemplos os nomes compostos hifenizados) escritos em uma língua fonte acompanhados de suas traduções na língua alvo. A utilidade de exemplos de sentenças paralelas alinhadas é inegavelmente grande, e as informações sobre as estruturas dessas sentenças e as correspondências existentes entre elas são extremamente relevantes para pesquisas em linguagem natural (MATSUMOTO et al. 2003 *apud* CASELI, 2004).

Nos últimos anos, têm sido propostos vários métodos para extrair, de forma automática, as correspondências estruturais, sintáticas e/ou lexicais dos textos alinhados. Essas correspondências são chamadas de regras de tradução (ou de transferência), e são usadas, em sistemas de TA, para traduzir (transferir) a representação de uma sentença na língua fonte em uma representação correspondente na língua alvo (BOSTRÖM, 2000 *apud* CASELI, 2004).

Duas etapas importantes para o processo de indução de regras de tradução a partir de textos paralelos alinhados sentencialmente incluem: a identificação de padrões e a geração de regras de tradução e filtragem e/ou ordenação das regras geradas.

A identificação de padrões pode ser realizada, por exemplo, por meio de reconhecimento de sequências repetidas de palavras em dois pares de exemplos, ou por meio de correspondências lexicais em um léxico bilíngue (alinhamento lexical). Já a geração de regras de tradução é realizada com base nos padrões ou alinhamentos definidos na etapa de identificação. Os padrões

são agrupados e generalizados (partes do padrão são substituídas por variáveis), considerando apenas a existência de similaridades, diferenças, ou ambas (similaridades e diferenças).

As regras de tradução podem ser compostas por informações mais complexas, como as representadas no formalismo mostrado na Figura 1 (para os pares inglês-hindi). Uma regra de tradução com esse formato possui as seguintes informações: informação de tipo, informação morfossintática, alinhamentos, restrições do lado fonte, restrições do lado alvo e restrições de ambos os lados. Para fins desta pesquisa são descritas apenas as seguintes informações: de tipo e morfossintática. A informação de tipo define o tipo de uma regra de tradução que geralmente corresponde ao tipo de um constituinte sintático. Por exemplo, as regras para sentenças são do tipo S, para sintagmas nominais (*noun phrases*) do tipo NP, etc. Já a informação morfossintática lista os componentes de uma regra (categorias lexicais, itens lexicais, etc.) tanto para a língua fonte quanto para a língua alvo (CARBONELL et al., 2000 *apud* CASELI, 2004)

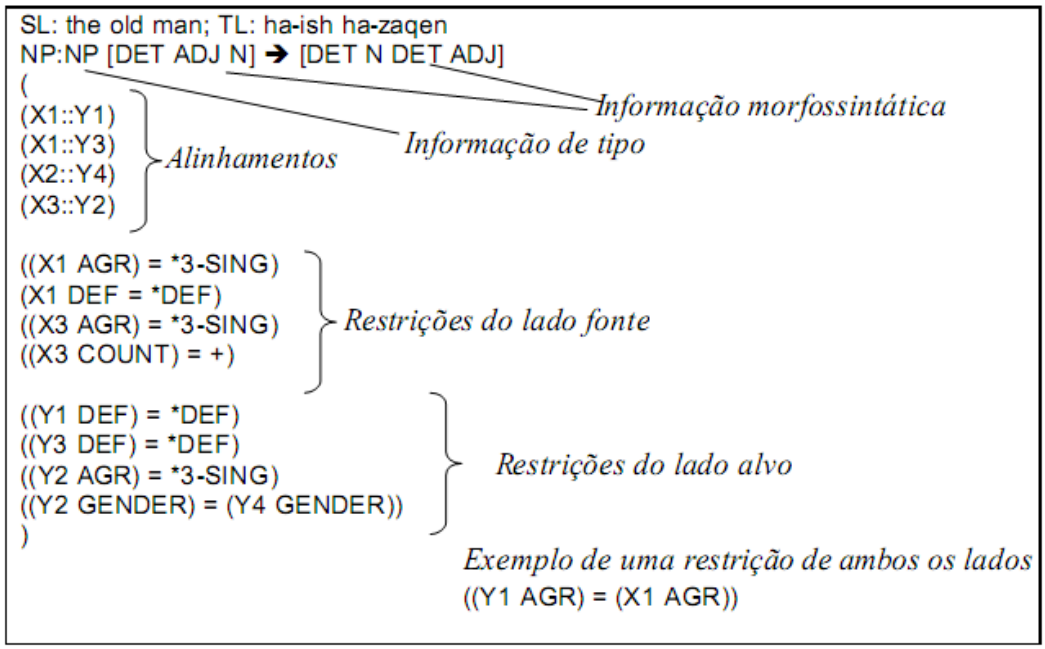

Figura 1 – Exemplo de formalismo de representação de regra de tradução (LAVIE et al., 2004 *apud* CASELI, 2004).

Nesta pesquisa, são utilizados exemplos de adjetivos e substantivos compostos hifenizados em inglês e suas traduções para o português, alinhados de acordo com a estrutura e

seus constituintes sintáticos, visando fornecer informação morfossintática para auxiliar a geração de regras de tradução em sistemas de tradução automática. Um dos exemplos é mostrado abaixo:

1. capacity-building
2. fortalecimento institucional e técnico

Nesse exemplo, a informação relativa à estrutura morfossintática para a geração de regras de tradução seria, em português:

Subs + Ing → Subs + Adj + Prep + Adj

Ou de acordo com o formalismo mostrado na Figura 1, em inglês:

N ING → N ADJ PRE ADJ

Exemplos de estruturas morfossintáticas semelhantes são analisados no capítulo referente aos resultados da análise das correspondências de tradução.

2.4 A análise de dados quantitativos

A diferença entre a análise de *corpus* qualitativa e quantitativa é que na primeira não se tenta atribuir frequências aos traços linguísticos identificados. Enquanto que na pesquisa quantitativa faz-se a classificação e contagem de traços linguísticos, na pesquisa qualitativa os dados são usados apenas como base para identificação e descrição de aspectos do uso na língua ou fornecer exemplos de um fenômeno particular (McENERY E WILSON, 2001).

Quando se aplica estatística à análise de dados linguísticos, a maioria das técnicas trabalha com a frequência de palavras. Em certos casos podem ser construídos modelos estatísticos complexos numa tentativa de explicar o que está sendo observado (KILGARRIFF, 2001; McENERY E WILSON, 2001).

As duas estatísticas comumente usadas para comparar a frequência de palavras entre dois corpora são o teste qui-quadrado e o *likelihood ratio* (ou razão de probabilidades logarítmicas). Segundo Dunning (1993), em casos nos quais o teste qui-quadrado funciona bem, os testes de razão de probabilidades logarítmicas (*likelihood ratio*) podem ser usados com aproximadamente a mesma eficiência.

Para a comparação das frequências do *corpus* de referência e do *corpus* de estudo utilizou-se a estatística *log-likelihood*. Selecionou-se a opção *log-likelihood* (ferramenta *Keywords*) que realiza o teste de razão de probabilidades logarítmicas. Consequentemente, as palavras cujas frequências relativas no *corpus* de estudo são significativamente maiores em relação ao *corpus* de referência, ao serem identificadas, passam a compor uma listagem específica de palavras, chamadas de palavras-chave (BERBER SARDINHA, 2004, p. 97).

Rayson, Berridge e Francis (2004) investigaram a confiabilidade dos testes estatísticos qui-quadrado e de razão de probabilidades logarítmicas em *corpora* de tamanhos diferentes para comparar a frequência de palavras. Para estender a aplicabilidade das comparações de frequência a valores esperados de 1 ou mais, o uso da razão de probabilidades logarítmicas foi considerado como sendo preferível em relação à estatística qui-quadrado.

De acordo com Dunning (1993), para variáveis aleatórias a distribuição de frequência subjacente a muitos testes estatísticos é a distribuição normal ou a distribuição qui-quadrado. Essa distribuição é simétrica em torno da média, o que em análise textual, significaria que as palavras de maior frequência se localizariam em torno da média. No entanto, dada a natureza não aleatória das palavras em um texto, isso não ocorre. Por isso, não se deve basear na suposição de uma distribuição normal quando se realiza análise estatística de um texto. Dunning (1993) sugere que a análise paramétrica baseada em distribuição binomial ou multinomial é uma melhor alternativa para textos menores.

Na distribuição binomial, a tarefa de contar palavras pode ser computada na forma de uma sequência de eventos binários similares ao arremesso de uma moeda (no qual, por exemplo, o aparecimento de *cara* é um sucesso e *coroa* é fracasso). A contagem pode ser feita comparando cada palavra em um texto com a palavra que está sendo contada (DUNNING, 1993). Por exemplo, em termos simples, a palavra a ser contada é *planning*, então cada palavra do texto será comparada a essa, e o número de sucessos (aparecimento de *planning*) indicará a probabilidade de encontrar essa palavra no *corpus*.

O teste qui-quadrado não é confiável quando a frequência esperada é menor que 5 e possivelmente superestima palavras de alta frequência quando se compara um *corpus* relativamente pequeno com um *corpus* muito maior. Já os métodos baseados nos testes de razão de probabilidades logarítmicas rendem bons resultados em amostras relativamente pequenas. Esses testes podem ser implementados de modo eficiente, e têm sido usados para a detecção de termos compostos, e para a determinação de termos específicos a certos domínios (DUNNING, 1993).

Dunning (1993) testou a eficácia dos métodos de probabilidades logarítmicas analisando uma amostra de 30.000 palavras de um texto obtido do Union Bank of Switzerland, com a intenção de encontrar pares de palavras que ocorressem próximas a outras, com uma frequência significativamente mais alta do que seria esperado com base nas frequências das palavras separadamente. Os resultados deveriam salientar as colocações mais comuns em inglês bem como as colocações peculiares à natureza financeira do texto analisado. Segundo Dunning (1993), a classificação com base nos testes de razão de probabilidades logarítmicas faz exatamente isso. Portanto, comparações semelhantes podem ser feitas entre um *corpus* grande de textos gerais e um *corpus* composto por textos de um domínio específico para produzir listas que consistem apenas em palavras e bigramas característicos de textos de domínios específicos.

Esta pesquisa utiliza técnicas quantitativas e qualitativas para examinar um *corpus* paralelo composto por textos de domínio restrito. O *corpus* geral utilizado, que servirá como *corpus* de referência, é o BNC (*British National Corpus*), composto por diferentes tipos de textos e domínios. No entanto, somente são usados os textos escritos desse grande *corpus* normativo. Foram geradas listas de palavras-chave usando o teste da razão de probabilidades logarítmicas (*likelihood ratio*), que está embutido no programa *WordSmith Tools*, para verificar as palavras em sobre-uso no *corpus* especializado (Agenda 21), em relação ao *corpus* geral (BNC).

2.5 Substantivos e adjetivos compostos hifenizados

De acordo com Carter e McCarthy (2006), sintagmas nominais em inglês (*noun phrases*) consistem em no mínimo um substantivo ou pronome, que age como o núcleo do sintagma

nominal. O núcleo pode ser acompanhado de elementos dependentes colocados antes e depois dele.

Quanto à função, os sintagmas nominais são expressões usadas para se referir a casos específicos ou classes gerais de pessoas e coisas. Eles funcionam tipicamente na oração como sujeitos, objetos, complementos e ocasionalmente como adjuntos.

Os sintagmas nominais podem ter núcleos simples, como em *my sister,* ou compostos, como em *window box*. No caso de núcleos compostos, Carter e McCarthy (2006) usam os termos substantivos compostos e adjetivos compostos.

Em inglês, os nomes compostos (*compounds*) consistem em um núcleo de substantivo colocado com outro elemento (na maioria das vezes um substantivo, mas pode ser também um adjetivo ou verbo) colocado antes desse núcleo em uma relação semântica e sintática muito próxima. O elemento inicial identifica na maioria das vezes um tipo de classe de entidades denotadas pelo substantivo final. Por exemplo, *video shop*, é um tipo de loja (shop); *orange juice*, é um tipo de suco (juice) (CARTER E MCCARTHY, 2006).

Os elementos em nomes compostos estão intimamente ligados uns aos outros sintaticamente e normalmente não podem ser interrompidos por outros elementos (por exemplo*, a motorway* ***petrol station***, e não *a* ***petrol*** *motorway* ***station***). Portanto, nomes compostos são considerados como núcleos simples em um sintagma nominal (noun phrase). Eles apresentam um padrão típico de ênfase, colocada sobre o primeiro elemento (por ex., **petrol** station).

A linha divisória entre substantivos compostos e os demais sintagmas nominais – que agem como pré-modificadores de núcleos de substantivos – nem sempre é clara. Porém, o padrão de ênfase (stress) para os nomes compostos com ênfase no primeiro elemento é geralmente uma indicação de que os substantivos são considerados como uma unidade 'institucionalizada', como em ***bus*** *stop* e ***safety*** *helmet*. A construção com o modificador do substantivo tem a ênfase no núcleo (elemento final), como em *government* ***report*** (CARTER E MCCARTHY, 2006).

A estrutura do nome composto é extremamente variada, apresentando os diferentes tipos de relacionamentos de significado que ele pode indicar. Ele pode ser usado para indicar o que alguém faz (*language teacher*), para que algo serve (*waste-paper basket, grindstone*), quais as qualidades de alguma coisa (*blackboard*), como alguma coisa funciona (*immersion heater*) quando algo acontece (*night frost*), onde alguma coisa está (*doormat*), de que algo é feito (*woodpile*), e assim por diante.

De acordo com Carter e Mccarthy (2006), em substantivos compostos existe uma série de relacionamentos semânticos possíveis entre o item pré-núcleo e o núcleo. Certos substantivos compostos são formados com –ing, apresentando os seguintes relacionamentos:

verbo + sujeito: *warning sign* (sign that warns)
objeto + verbo: *risk-taking* (that takes risks)

Muitos adjetivos compostos terminam em um adjetivo (por ex. air-sick) ou uma forma –ed ou –ing adjetiva, como *heart-breaking, white-washed*. Os principais relacionamentos entre as partes de adjetivos compostos são os seguintes:

- Objeto + -ing:

English-speaking (speaks English)

confidence-boosting (boosts confidence),

heart-breaking (breaks hearts)

- Complemento de verbo + -ing:

far-reaching (reaches far)

- sujeito + complemento de predicativo: *top-heavy* (the top is heavy)
- (A is B) comparativo: *paper-thin* (as thin as paper)
- (as B as A) adjetivo + complemento:

fat-free (is free of fat)

user-friendly (friendly to the user)

- adjetivo + núcleo adjetivo: *royal-blue, light-green, bitter-sweet*.

Alguns adjetivos compostos são formados pela adição da inflexão -ed a um substantivo:

right-angled (formed from right angle), *left-handed* (formed from left hand).

O uso de hífens em nomes compostos e palavras complexas envolve várias regras diferentes, e segundo Carter e McCarthy (2006) essa prática está mudando, sendo empregados

poucos hífens no uso contemporâneo. Quando um nome composto modifica um núcleo, normalmente é inserido um hífen para indicar quais palavras são compostas (por ex., *a well-known entertainer*). Em adjetivos numericamente modificados, todos os elementos modificadores são hifenizados, mas essas formas são usadas apenas atributivamente (e.g. *an eighteen-year-old girl, a twenty-ton truck, a twenty-four-hour flight*) (CARTER E MCCARTHY, 2006).

A tradução de substantivos e adjetivos compostos (englobados aqui no termo 'nomes compostos') para especificar regras de tradução será o foco desta pesquisa, que dará ênfase às formas hifenizadas que coocorrem com algumas palavras-chave, conforme será explicado em maiores detalhes no capítulo de metodologia.

2.6 Características de Documentos Oficiais

Thunes (1998) em sua análise empírica considerou documentos legais e descrições técnicas como 'textos de domínio restrito'. A pesquisadora utiliza o termo "correspondências de tradução", que significa como se dá a relação entre unidades lexicais e as construções sintáticas de dois sistemas de linguagem. No entanto, ela analisa a tradução de sequências de palavras (strings) mais longas e não apenas nomes compostos, como é feito neste trabalho.

Thunes (1998) verificou que textos de domínio restrito, como acordos oficiais, são escritos em estilo formal e impessoal, com o uso freqüente de sentenças longas e complexas. Outras características são a coordenação complexa e ocorrência de numerosas construções não-finitas.

Biber (1988) analisou diversos documentos oficiais que incluem vários tipos de discursos, incluindo relatórios de governos, documentos legais e tratados, relatórios de negócios, entre outros. Esse gênero é fortemente restrito na sua forma linguística, e geralmente não há um autor que possa ser identificado. Biber (1988) ao comparar diferentes gêneros de discurso oral e escrito, constatou que documentos oficiais possuem as seguintes características:

- caráter altamente informacional, apresentando uma frequência muito alta de substantivos e preposições;

- poucos verbos em relação aos outros gêneros discursivos, e os existentes estão no infinitivo ou em construções passivas;
- uso do tempo presente (geralmente);
- muitos adjetivos atributivos;
- muitas frases longas e uma seleção cuidadosa do vocabulário, resultando em uma razão forma/ocorrência alta (type/token ratio) em relação aos outros gêneros;
- discurso nominal, e descritivo ou argumentativo.

Esta pesquisa procurará identificar no *corpus* de estudo (Agenda 21), que consiste em textos que compõem um documento oficial, algumas das características típicas desse gênero ou domínio, conforme Biber (1988).

3 Metodologia

Os *corpora* utilizados para esta pesquisa consistem em textos originais em inglês e sua tradução para o português, que são chamados de *corpora* paralelos, ou mais especificamente, *corpora* de tradução.

Em primeiro lugar, a coleta do *corpus* foi feita considerando apenas um gênero ou domínio: documentos oficiais. Os textos compõem o documento da Agenda 21, uma política pública resultante de um acordo oficial endossado por 179 países na Conferência Eco-92 (1992 United Nations Conference on Environment and Development – UNCED), também conhecida como *The Earth Summit.*

O *corpus* paralelo foi extraído do *site* internacional do Programa para o Meio Ambiente das Nações Unidas (United Nations Environment Programme – UNEP) (textos em inglês) e sua tradução para o português foi coletada do *site* da Secretaria de Estado para o Meio Ambiente de São Paulo, disponíveis nos seguintes URLs:

- http://www.unep.org/Documents.Multilingual/Default.asp?DocumentID=52
- http://www.ambiente.sp.gov.br/agenda21/indice.htm

O *corpus* inclui todos os capítulos do documento (quarenta capítulos), que totalizam 147.365 palavras no documento em inglês, e 170.590 palavras na tradução para o português.

Podemos considerar o *corpus* de estudo como um *corpus* de tamanho médio, ou seja, que possui entre 80 a 250 mil palavras. O tamanho mínimo do *corpus* especializado deve ser de 91.161 ocorrências (BERBER SARDINHA, 2004).

Os capítulos em inglês foram salvos separadamente e colocados numa pasta denominada Ag21english. O mesmo foi feito com os capítulos em português, que foram colocados numa pasta chamada de Ag21port.

O *corpus* de referência usado foi o BNC (*British National Corpus*), cujos textos escritos apresentam um total de 90.748.880 ocorrências. A lista de palavras já pronta desse *corpus* foi obtida no endereço http://www2.lael.pucsp.br/corpora/bi/listas/bnc/. A lista de palavras foi salva na pasta BNCWritt.

A análise quantitativa foi realizada através do programa *Wordsmith Tools*, (Scott, 1996 – versão 3.00.00) usando as ferramentas *Wordlist, Keywords* e *Concord.*

3.1 Análise das correspondências de tradução

Essa análise foi realizada a partir das evidências fornecidas pelo *corpus*, através do uso das ferramentas *Keywords* e *Concord*, do programa *WordSmith Tools*. Na análise das correspondências de tradução de substantivos e adjetivos compostos hifenizados foram observados certos padrões, que podem ser usados para auxiliar a geração de regras para a tradução dos substantivos e adjetivos compostos hifenizados.

Primeiramente foi utilizada a ferramenta *Keywords*, que contrasta uma lista de palavras (ou mais de uma) de um *corpus* de estudo com uma lista de palavras de um *corpus* de referência (aqui, o BNC), produzindo uma terceira lista contendo somente as palavras-chave do *corpus* de estudo (BERBER SARDINHA, 2005).

Conforme recomendado por Berber Sardinha (2004), o tamanho do *corpus* de referência influencia a quantidade de palavras-chave obtidas. Este autor recomenda que os *corpora* de referência sejam 2, 3 a 5 vezes o tamanho do *corpus* de estudo. *Corpora* de referência com essas dimensões retornam significativamente mais palavras-chave do que corpora de tamanhos menores. Outra observação feita por esse autor ressalta que o *corpus* de referência não deve conter o *corpus* de estudo, pois isso diminui as chances de determinadas palavras serem chave, ou seja, diminui a propensão à chavicidade.

A seguir, utilizou-se a ferramenta *Concord* que produz concordâncias ou listagens das ocorrências de um item específico acompanhado do texto ao seu redor (o cotexto). No *WordSmith Tools*, o *Concord* pode ser usado separadamente, para concordâncias avulsas, ou em conjunto com as ferramentas *Wordlist* e *Keywords*, chamado a partir desses programas. Essa ferramenta foi aplicada para buscar todas as ocorrências mais frequência da *Wordlist* Ag21.lst. Aqueles itens lexicais que apresentavam frequência muito alta tiveram seus primeiros colocados à esquerda (L1) examinados. As ocorrências que apresentavam o maior número de substantivos ou adjetivos compostos hifenizados em L1 foram escolhidas. São elas os substantivos gerais *systems,*

programmes e *activities*. Essas ocorrências e suas colocações com substantivos e adjetivos compostos hifenizados foram então analisadas, para descobrir as correspondências de tradução dos mesmos para o português. Aqui, o número de ocorrências desses substantivos e adjetivos compostos hifenizados foi verificado, considerando-os como uma unidade. Com o propósito de verificar as correspondências de tradução, foi feito o alinhamento das sentenças em inglês e português de modo manual, no programa *Word*, da suíte Microsoft Office. Não foi necessário fazer o alinhamento do *corpus* inteiro, uma vez que foram selecionadas apenas algumas sentenças para análise. Tais sentenças continham os nomes compostos hifenizados em inglês e suas traduções para o português, que foram localizados (pelo uso da ferramenta 'Localizar') nos quarenta arquivos em formato 'Text' (português e inglês) que compuseram o *corpus.* O critério de escolha para a seleção desses compostos hifenizados foi baseado, preferivelmente, na existência do segundo elemento terminado em –ed particípio passado ou –ing particípio presente. No entanto, foram incluídos alguns casos que não seguiam esse critério, principalmente os que *não* apresentavam uma tradução direta (palavra-por-palavra).

A análise das correspondências de tradução dos substantivos e adjetivos compostos hifenizados foi feita com base na comparação da estrutura morfossintática em inglês e sua tradução para o português.

3.2 Análise das características do *corpus* de estudo

Esta etapa parte de certas características pré-definidas para documentos oficiais, como as apresentadas por Biber na seção 2.4, e utiliza o *corpus* de estudo para quantificar e validar tais características.

Com essa finalidade, em primeiro lugar foi utilizada a ferramenta *Wordlist*, que como o nome indica, cria listas de palavras, ordenando-as por frequência e alfabeticamente. A informação sobre a frequência é muito útil para identificar as características de um texto ou gênero/domínio. Essa ferramenta também gera informações estatísticas como a razão tipo/ocorrência (*type/token ratio*), número de sentenças e comprimento médio das sentenças (*average sentence length*).

Foram geradas três janelas, uma com as informações estatísticas mencionadas acima, uma com a frequência de palavras listadas alfabeticamente, e uma com as palavras listadas em ordem decrescente de frequência. A lista de palavras em inglês – Ag21.lst – foi salva. Foi feito o mesmo procedimento para os textos da pasta Ag21port e depois para a pasta BNCWritt, salvando as listas de palavras como Ag21port.lst e BNCWritt.lst (essa pasta não continha a Ag21.lst), respectivamente.

Posteriormente, foi feita a etiquetagem morfossintática, por meio de anotação manual (na lista de palavras em fomato 'Text'), para as ocorrências da lista em inglês, que traz as palavras em ordem decrescente de frequência. Nessa etapa do trabalho, os adjetivos e substantivos compostos hifenizados foram considerados separadamente, ou seja, os hífens são considerados como separadores de palavra.

Os tipos – também chamados por Biber (2004) de 'formas' – foram classificados em quinze (15) classes gramaticais: substantivo, preposição, adjetivo, verbo, verbo auxiliar, modal, advérbio, pronome, conjunção[4], artigo, acrônimo, sufixo, elemento de composição[5], outros idiomas – compostos de palavras de outros idiomas – e erro, que considera os possíveis erros relacionados à limpeza[6] e contagem das ocorrências do *corpus* de estudo.

Essa parte da pesquisa foi realizada cuidadosamente, principalmente pelo fato de que muitos tipos podem pertencer a mais de uma classe gramatical. Por exemplo, a maioria dos tipos que termina em –ed ou –ing pode ser considerada como adjetivo ou verbo ('planned', 'damaging'). Além disso, outros tipos também podem pertencer a uma ou mais classes gramaticais: 'increase' pode ser tanto substantivo como verbo; 'either' pode ser conjunção ou pronome, etc. Portanto, foi preciso consultar cada um desses tipos a partir da ferramenta *Concord* do programa Wordsmith, e então computar, por exemplo, quantas ocorrências correspondem à classe dos adjetivos e quantas são consideradas verbos na frase em questão.

A partir da etiquetagem dos tipos, foi criada uma planilha no programa Excel, com quinze colunas correspondentes a cada uma das classes gramaticais citadas acima. A partir disso,

---

[4] Essa palavra terá relevância na análise feita mais adiante, significando "Palavra invariável que liga duas orações ou dois termos semelhantes da mesma oração" (Dicionário Aurélio Digital, Versão 5.11ª, 2004).

[5] Forma lingüística que só ocorre em vocábulos compostos ou derivados (*aerofagia*, *fagícola*, *aglutinina*), especialmente aqueles da terminologia científica, e é constituída ou por um radical isolado (*-i*), ou por um radical seguido de vogal de ligação (*flum(i)-*, *pneumat(o)-*), ou por um radical associado a um prefixo (*acefal(o)-*) ou a um sufixo (*-logia*) (Dicionário Aurélio Digital, Versão 5.11ª, 2004).

[6] Erros relacionados ao aparecimento de códigos de formatação (em linguagem HTML) nos capítulos do *corpus*, salvos em formato 'txt'.

cada tipo foi classificado para uma das classes/colunas. Por exemplo, o tipo 'development' ocorreu 1.403 vezes, e esse número de ocorrências foi colocado na coluna de substantivos. O tipo 'of' ocorreu 7.195 vezes e esse número foi colocado na coluna de preposições. Acabada a classificação de todos os tipos, foi feito o somatório para cada classe gramatical (ou seja, para cada coluna da planilha). Em seguida, foi realizado o somatório final, incluindo os subtotais de cada classe. Esse somatório deveria ser igual ao número total de ocorrências fornecido pelo programa *WordSmith Tools*.

Essa abordagem foi utilizada para atingir um dos objetivos da pesquisa: verificar as principais características de documentos oficiais, apresentadas por Biber (1988).

4 Resultados

4.1 Resultados Gerais

A estatística geral dos textos em inglês é mostrada na Tabela 1, a seguir.

Tabela 1 – Dados estatísticos resultantes da lista de palavras AG21.lst.

Ocorrências	**147.365**
Tipos	**5.427**
Razão tipo/ocorrência	3,55
Razão tipo/ocorrência padronizada	37,68
Comprimento médio das palavras	5,87
Sentenças	3.383
Comprimento das sentenças	22,21
Comprimento padronizado das sentenças	28,13
Parágrafos	3.875
Comprimento dos parágrafos	39,11
Comprimento padronizado dos parágrafos	143,77

Os dados estatísticos mostram o total de ocorrências no *corpus,* devendo-se salientar novamente que todos os substantivos e adjetivos compostos hifenizados foram considerados como duas ocorrências distintas, ou seja, o hífen é considerado separador de palavras. Deste modo, temos um total de 147.365 ocorrências, e 5.427 tipos. Já na tradução para o português, as ocorrências somam 170.581 palavras, e ocorrem 7.298 tipos.

Os tipos mais frequentes são mostrados na Tabela 2. A classe gramatical dos tipos será mostrada na última coluna.

Tabela 2 – Tipos mais frequentes da Ag21.lst.

PALAVRA	**FREQUÊNCIA**	**%**	**CLASSE GRAMATICAL**
AND	11.153	7,57	Conjunção = cj
THE	8.066	5,47	Artigo = at
OF	7.191	4,88	Preposição = pp
TO	4.328	2,94	Preposição = pp
IN	2.974	2,02	Preposição = pp
FOR	2.141	1,45	Preposição = pp
DEVELOPMENT	1.402	0,95	Substantivo = sb
SHOULD	1.401	0,95	Modal = md
ON	1.224	0,83	Preposição = pp
WITH	1.219	0,83	Preposição = pp
AS	1.157	0,79	Preposição = pp
BE	1.028	0,70	Verbo Auxiliar = va
A	1,401	0,69	Artigo = at
INTERNATIONAL	1.000	0,68	Adjetivo =aj
COUNTRIES	826	0,56	Substantivo =sb
APPROPRIATE	818	0,53	Adjetivo = aj
ARE	787	0,51	Verbo auxiliar =va
NATIONAL	782	0,51	Adjetivo = aj
THAT	776	0,50	Conjunção = cj
ORGANIZATIONS	735	0,50	Substantivo = sb
BY	710	0,48	Preposição = pp
MANAGEMENT	712	0,48	Substantivo = sb
INCLUDING	707	0,48	Preposição =pp
PROGRAMMES	**704**	**0,48**	Substantivo = sb
ACTIVITIES	**692**	**0,47**	Substantivo = sb
RESOURCES	681	0,46	Substantivo = sb
:	:	:	:
SYSTEMS	**329**	**0,22**	Substantivo = sb

As palavras que compõem a lista de palavras-chave do *corpus* de estudo, tendo como *corpus* de referência o BNC são mostradas no Anexo A. Algumas dessas palavras são mostradas na Tabela 3.

Tabela 3 – Palavras-chave da Ag21.lst (continua)

PALAVRA-CHAVE	FREQ.AG21.LST	FREQ.BNC.LST	CHAVICIDADE
AND	11.187	32.815	8.508,6
DEVELOPMENT	1.403	31.114	6.458,0
SUSTAINABLE	613	654	6.072,6
ORGANIZATIONS	736	2.799	5.789,0
INTERNATIONAL	1.000	21.761	4.636,2
APPROPRIATE	818	10.863	4.553,0
PROGRAMMES	704	6.212	4457,8
GOVERNMENTS	616	4.570	4.098,7
COUNTRIES	826	16.230	3.987,5
RESOURCES	681	9.930	3.669,9
ENVIRONMENTAL	633	7.968	3.586,4
ACTIVITIES	692	11.316	3.579,6
SHOULD	1.401	96.455	3.536,4
COOPERATION	416	1.248	3.442,8
REGIONAL	597	7.425	3.395,8
IMPLEMENTATION	450	2.799	3.139,9
DEVELOPING	527	5.972	3.089,6
MANAGEMENT	712	20.934	2.897,9
INCLUDING	707	23.694	2.897,9
NATIONAL	782	36.493	2.515,8
CONCESSIONAL	203	25	2.433,3
NATIONS	397	4.028	2.409,5
ENVIRONMENTALLY	264	665	2.262,9
ENVIRONMENT	509	12.417	2.249,0
PROMOTE	343	3.047	2.166,8
RELEVANT	407	7.677	1.995,1
WASTES	227	617	1.916,6

Tabela 3 – Palavras-chave da Ag21.lst (Continuação)

GOVERNAMENTAL	231	900	1.803,6
INTEGRATED	197	2.632	1.093,5
LAND	364	20.231	1,055,1
IMPLEMENTING	147	832	1.050,8
RELATED	*279*	*9.749*	*1.045,2*
AGENDA	175	1.948	1.031,4
TECHNOLOGICAL	166	1.644	1.014,8
CONSERVATION	204	3.873	997,2
SYSTEMS	329	17.011	996,4
.	.	.	.

A razão tipo/ocorrência (ou forma/item, conforme utilizado por Berber Sardinha, 2004) mostra quantas palavras diferentes existem no *corpus*, indicando a riqueza lexical do texto. A razão tipo/ocorrência (TO) é obtida dividindo-se o total de tipos pelo total de ocorrências, levando em conta todas as palavras do *corpus* selecionado. No Wordlist, transforma-se esse valor em porcentagem, dividindo-se a razão tipo/ocorrência por cem. O procedimento para o cálculo da razão tipo/ocorrência padronizada consiste em segmentar o *corpus* em partes de 10.000 palavras, calcular a razão tipo/ocorrência para cada segmento e depois tirar a média. Isso impede que o efeito da repetição seja tão marcado quanto o observado quando o total de tipos e ocorrências do *corpus* é considerado, uma vez que a proporção não sofre tanto o impacto da repetição das palavras gramaticais.

A diferença entre os valores de TO simples e padronizado ocorre devido ao fato de o *corpus* inteiro, por ser maior, dá mais espaço para repetições, por isso o valor TO é mais baixo. No cálculo padronizado, impediu-se que se levasse em conta a repetição de palavras ocorridas no outro trecho, resultando em um valor médio mais alto (BERBER SARDINHA, 2004).

Para o *corpus* de estudo usado nesta pesquisa, a razão tipo/ocorrência para o *corpus* em inglês é 3,55, e a razão tipo/ocorrência padronizada é 37,68. Para o *corpus* em português, as razões apresentadas são 4,28 e 38,62, respectivamente. Em um *corpus* de referência grande,

como o NILC[7] a razão TO é 0,80, e a TO padronizada, 46,63. Portanto, a diferença entre razão tipo/ocorrência padronizada entre o *corpus* de estudo e o NILC não é muito grande. Podemos afirmar então que apesar do número de repetições, o *corpus* da Agenda 21 possui considerável variedade vocabular.

Como seria esperado, primeiramente aparece uma maior frequência de itens gramaticais. A alta frequência do item "*should*'' mostra o aspecto altamente prescritivo dos textos que compõem esse documento oficial, como será exemplificado mais adiante.

Os itens lexicais mais frequência estão relacionados à área de gestão ambiental (*development, international, national, countries, organization, management, programmes, activities, systems*). Os itens lexicais – *programmes, activities e systems* – podem ser considerados como substantivos gerais (*general nouns)* e foram escolhidos devido a sua ocorrência com nomes compostos hifenizados. Esses itens e seus colocados foram considerados interessantes para fins de análise das correspondências de tradução.

Os itens *programmes, activities* e *systems* são polissêmicos, mas neste *corpus* eles são usados apenas com o seguinte sentido:

- *A **programme** of actions or events is a series of actions or events that are planned to be done.*
- *The **activities** of a group are the things that they do in order to achieve their aims.*
- *A **system** is a way of working, organizing, or doing something which follows a fixed plan or set of rules (COLLINS COBUILD DICTIONARY, 2002).*

Cada adjetivo ou substantivo composto hifenizado que coocorre com os substantivos *programmes, activities e systems* serve para determinar a espécie de cada programa, atividade ou sistema que está sendo citado no documento sob análise.

O número total de ocorrências para cada adjetivo ou substantivo composto hifenizado ***em inglês*** (aqui, considerado como uma unidade) é mostrado antes da análise das correspondências de tradução. Deve-se salientar que alguns dos substantivos e adjetivos

[7] Corpus do Núcleo Interinstitucional de Linguística Computacional da USP/São Carlos, formado na maior parte por textos jornalísticos. No entanto julgou-se ser possível a comparação da razão TO desse corpus com o corpus de estudo.

compostos mostrados abaixo também aparecem no *corpus* de estudo *sem hífen*, no entanto, essas ocorrências não foram consideradas, por não serem frequência.

4.2 Resultados da análise das correspondências de tradução

Esta pesquisa identificou quarenta e um exemplos de substantivos e adjetivos compostos hifenizados, cujas correspondências de tradução foram analisadas. Os exemplos foram agrupados dentro de onze grupos denominados 'padrões', de acordo com a estrutura morfossintática observada na *tradução para o português.*

PRIMEIRO PADRÃO

As correspondências de tradução para substantivos e adjetivos compostos hifenizados consistem no uso de apenas um tipo (no caso, um adjetivo) na tradução para o português. Nos seguintes compostos hifenizados o segundo elemento em inglês – geralmente formado por –ed particípio passado – é omitido completamente. No caso em que os dois primeiros elementos consistem em adjetivo, a tradução corresponde a um único adjetivo. Isso pode ser verificado através dos exemplos mostrados abaixo:

1. **land-based** (13 ocorrências); **sea-based** (4 ocorrências)

1a. Regularly exchanging information on marine degradation caused by **land-based** *and* **sea-based** *activities*…
1b. Intercambiar regularmente informações sobre a degradação marinha causada tanto por *atividades* **terrestres** como **marítimas**...

Em 1a, os nomes compostos ***land-based*** *and* ***sea-based*** (subst + -ed part. passado) e (subst + -ed part. passado) são formas combinadas que significam '*activities that are based on land or sea'*. Em 1b a forma (–based) é completamente omitida, a tradução para o português

consiste em atividades **terrestres (adj)** como **marítimas (adj),** *ou seja,* nos dois casos utilizou-se apenas um adjetivo, que corresponde ao substantivo do composto em inglês, ou seja, houve omissão e mudança de classe gramatical durante o processo de tradução.

2. **cross-sectoral** (8 ocorrências)

2a. The specific requirements for the implementation of the sectoral and **cross-sectoral** *programmes* included in Agenda 21…
2b. As exigências específicas para a implementação dos programas **intersetoriais** incluídos na Agenda 21…

Em 2a, o adjetivo composto **cross-sectoral** (adj_1 + adj_2), que indica '*programmes extended to several sectors',* o que corresponde em português (2b) à forma aglutinada **intersetoriais (adj)** (considerado como um só tipo).

3. **action-oriented** (2 ocorrências)

3a. **Action-oriented** *activities* of relevance to the above objectives, such as poverty eradication…
3b. …deve-se ver com especial atenção as atividades **práticas** relacionadas a esses objetivos, como as de erradicação da pobreza...

Em 3a, o substantivo composto **action-oriented** (subst + -ed particípio passado) significa '*activities oriented to action'*, e a tradução para o português apresenta-se como (atividades) **práticas (adj).** O termo 'relacionadas' se refere à 'relevance to the above objectives'.

SEGUNDO PADRÃO

O composto em inglês, formado por dois substantivos, ou substantivo + -ed particípio passado, ou substantivo + -ing particípio presente, também é traduzido por apenas um

substantivo, no entanto este é precedido de uma preposição, que aqui não é considerada como parte da estrutura morfossintática do termo traduzido. Isso ocorre nos seguintes exemplos:

4. **awareness-creation** (1 ocorrência)

4a. **Awareness-creation** *programmes*, including mobilizing commitment and support at all levels…
4b. ...*programas* de **conscientização**, com a mobilização de compromisso e apoio em todos os níveis...

Em 4a, o substantivo composto **awareness-creation** ($subst_1$ + $subst_2$), que indica '*programmes that create awareness'*, corresponde em 4b a programas (de) **conscientização (subst)**. Com isso, o termo '*awareness*' que em inglês significa 'consciência' em português incorporou o sufixo causativo –ização, significando o ato ou efeito de conscientizar.

5. **awareness-building** (2 ocorrências)

5a. …level with preference given to local responsibility and control over **awareness-building** *activities.*
5b. ...dar preferência para a responsabilidade e controle locais sobre as atividades de **conscientização**.

Em 5a, o substantivo composto **awareness-building** (subst + -ing particípio presente) significa '*activities that build awareness'*, correspondendo a atividades (de) **conscientização**. Nesse exemplo o termo 'awareness' também foi traduzido do mesmo modo como no exemplo anterior: '**conscientização**' **(substantivo).**

6. **field-level** (2 ocorrências)

6a.…Coordinator of UNDP needs to be strengthened in order to coordinate the **field-level** *activities,…*

6b. ...coordenador residente do PNUD a fim de coordenar as atividades **de campo**...

Em 6a, o substantivo composto **field-level** ($subst_1$ + $subst_2$), corresponde em 6b a atividades (de) **campo (subst),** omitindo o substantivo 'level' na tradução para o português.

7. **market-oriented** (5 ocorrências)

7a. …in particular the economies in transition from planned to **market-oriented…** *systems*…
7b. …Em particular, as economias nacionais em transição de sistemas de planejamento para sistemas de **mercado…**

Em 7a, o substantivo composto **market-oriented** (subst + -ed particípio passado) significa '*systems oriented to the market'*, corresponde em português a sistemas (de) **mercado (subst).** Neste exemplo, mais uma vez, omite-se o segundo elemento, formado por –ed particípio passado.

TERCEIRO PADRÃO

Há casos em que o primeiro elemento do nome composto hifenizado é advérbio e o segundo elemento é –ed particípio passado. Esses casos têm como correspondente em português dois elementos, sendo que o segundo é um advérbio, sempre acompanhado de particípio. A tradução é direta, e não existe mudança de classe gramatical no processo de tradução.

8. **above-mentioned** (6 ocorrências)

8a. UNESCO might take the lead in implementing the **above-mentioned** *activities*…
8b. A UNESCO poderia dirigir a implementação das atividades **acima mencionadas**,...

Em 8a o substantivo composto **above-mentioned** (adv + -ed particípio passado), que significa '*activities mentioned above',* corresponde a (atividades) **acima mencionadas (adv + part).**

QUARTO PADRÃO

O nome composto hifenizado consiste em uma preposição ou advérbio mais substantivo, cuja correspondência é um advérbio + adjetivo. Há ou não mudança de classe gramatical.

9. **Off-farm** (4 ocorrências)

9a. Initiate and maintain on-farm and **off-farm** *programmes* to collect and record indigenous knowledge…
9b. Iniciar e manter programas agrícolas e **não-agrícolas** para coletar e registrar os conhecimentos autóctones.

Neste exemplo o composto **off-farm** (prep + subst), que indica '*programmes (that take place) off-farm'*, correspondendo em português a programas **não-agrícolas,** ou seja**, (adv + adj)**. Do mesmo modo, **on-farm** (prep + subst) corresponde a agrícolas **(adv + adj).**

10. **non-farm** (6 ocorrências)

10a. …wildlife utilization, aquaculture and fisheries, **non-farm** *activities*, such as…
10b. ...fauna silvestres, aqüicultura e piscicultura, atividades **não-agrícolas** como...

No substantivo composto **non-farm** (adv + subst) a tradução correspondente é igual à do termo 'off-farm', mostrado no exemplo acima, ou seja, **não-agrícola (adv + adj).**

QUINTO PADRÃO

Os substantivos compostos hifenizados são formados por um substantivo ou adjetivo como primeiro elemento, tendo a forma –ing ou substantivo como segundo elemento. Nesses casos a tradução se dá pelo uso de dois substantivos, ou um substantivo e um adjetivo, ligados por uma preposição, conforme os exemplos a seguir.

11. **water-use** (8 ocorrências)

11a. To have established efficient **water-use** *programmes* to attain sustainable resource utilization patterns…
11b. Ter estabelecido programas eficientes de **uso de água** para alcançar padrões sustentáveis de utilização dos recursos.

Em 11a o substantivo composto **water-use** ($subst_1$ + $subst_2$) que indica '*programmes that use water',* corresponde em 11b a **uso de água ($subst_1$ + prep + $subst_2$).**

12. **land- and water-management** (3 e 1 ocorrências, respectivamente)

12a. …administrators and officers at all levels involved in **land- and water-management** *programmes.*
12b. ...administradores e funcionários de todas as categorias envolvidos em programas de **manejo de terra e água.**

Em 12a os substantivos compostos **land-** and **water-management** ($subst_1$ + $subst_2$), que correspondem a '*programmes that manage land (use) e programmes that manage water (use)'*, são traduzidos por **manejo de terra** e **(manejo de) água ($subst_1$ + prep + $subst_2$).**

13. **water-quality** (25 ocorrências)

13a. International **water-quality** *programmes*, such as GEMS/WATER, should be oriented towards the water-quality…
13b. Programas internacionais de **qualidade de água** como o GEMS/WATER devem ser orientados para o estudo da qualidade da água...

Em 13a, o substantivo composto **water-quality**, ($subst_1$ + $subst_2$) que significa '*programmes (that improve) the quality of water',* corresponde em português à **qualidade de água ($subst_1$ + prep + $subst_2$).**

14. **data-collection** (4 ocorrências)

14a. …United Nations system and relevant international organizations, **data-collection** *activities*,…
14b. ...do sistema das Nações Unidas e das organizações internacionais pertinentes, é preciso reforçar as atividades de **coleta de dados**,...

Em 14a o substantivo composto **data-colletion** ($subst_1$ + $subst_2$), que significa '*activities that collect data'*, é traduzido pelo correspondente **coleta de dados ($subst_1$ + prep + $subst_2$).**

15. **income-generating** (3 ocorrências)

15a.…combating poverty by strengthening employment and **income-generating** *programmes.*
15b ...erradicação da pobreza por meio do fortalecimento dos programas de emprego e **geradores de renda.**

Em 15a o substantivo composto **income-generating** (subst + -ing particípio presente), que corresponde a '*programmes that generate income',* é traduzido em 15b por **geradores de renda**, ou seja, (**adj + prep + subst).**

16. **cross-breeding** (1 ocorrência)

16a. (…) avoiding the risk of their being replaced by breed substitution or **cross-breeding** *programmes…*
16b. (...) e evitar o risco de que sejam substituídas por outras raças ou por programas de **cruzamento de raças.**

Em 16a, o adjetivo composto **cross-breeding** (adj + -ing part. pres.) significa '*breeds that are crossed'*, e em 16b foi traduzido diretamente como **cruzamento de raças***,* que consiste em (**$subst_1$ + prep + $subst_2$**).

17. **data- and statistics-gathering** (1 ocorrência)

17a. Coordinate existing ***data- and statistics-gathering*** *systems* relevant to developmental and environmental issues…
17b. Coordenar os sistemas atuais de **coleta de dados e estatísticas** pertinentes às questões de meio ambiente e desenvolvimento,...

Em 17a, o substantivo composto **data- and statistics-gathering** (subst + –ing particípio presente) indica '*systems that gather data and statistics'*, e em 17b as correspondências são dadas por **coleta de dados** e **(coleta de) estatísticas***,* ambas formadas por **$subst_1$ + prep + $subst_2$.**

SEXTO PADRÃO

Nesses casos o substantivo composto hifenizado em inglês é formado por: dois substantivos diferentes, ligados ou não por preposição, ou substantivo + -ing particípio presente, ou prefixo + substantivo, a tradução geralmente corresponde ao uso de dois substantivos (ou adjetivo + substantivo), que são ligados por preposição + artigo.

18. **life-support** (5 ocorrências)

18a.…guidelines relating to science and technology in which the integrity of **life-support** *systems…*
18b. ...e diretrizes relativos à ciência e tecnologia nos quais se leve em conta amplamente a integridade dos sistemas de **sustentação da vida**...

Em 18a, o substantivo composto hifenizado **life-support** ($subst_1$ + $subst_2$) corresponde em 18b a **sustentação da vida ($subst_1$ + [prep + art] + $subst_2$).**

19. **water-supply** (27 ocorrências)

19a. …local water associations and water committees to manage community **water-supply** *systems*…
19b. … associações e comitês de água locais para que gerenciem os sistemas de **abastecimento da comunidade**....

Em 19a, o substantivo composto **water-supply** ($subst_1$ + $subst_2$), *'systems thaat supply water',* corresponde a **abastecimento da comunidade ($subst_1$ + [prep + art] + $subst_2$)**, omitindo-se o termo 'água'.

20. **land-use** (37 ocorrências)

20a…environmentally sound, socially acceptable, fair and economically feasible **land-use** *systems*.
20b. …por meio da introdução de sistemas de **uso da terra** saudáveis, socialmente aceitáveis, justos e economicamente viáveis...

Em 20a, o substantivo composto **land-use** ($subst_1$ + $subst_2$) que significa **'***systems that use the land'*, corresponde a **uso da terra ($subst_1$ + [prep + art] + $subst_2$).**

21. **resource-use** (1 ocorrência)

21a. …to desertification and drought, current livelihood and **resource-use** *systems*…
21b. …à desertificação e à seca os sistemas vigentes de subsistência e **utilização dos recursos**…

Em 21a, o substantivo composto **resource-use** ($subst_1$ + $subst_2$), *'systems that use resource'* corresponde em português a **utilização dos recursos ($subst_1$ + [prep + art] + $subst_2$).**

22. **life-supporting** (2 ocorrências)

22a….examining in particular the capacities of global and regional **life-supporting** *systems…*
22b. examinando, em particular, a capacidade dos sistemas de **sustentação da vida** mundiais...

Em 22a, o substantivo composto **life-supporting** (subst + -ing particípio presente) corresponde a **sustentação da vida ($subst_1$ + [prep + art] + $subst_2$),** como no exemplo mostrado acima.

23. **right-to-know** (2 ocorrências)

23a. Adopt, on a voluntary basis, community **right-to-know** *programmes* based on international guidelines…
23b. Adotar a título voluntário programas reconhecendo o **direito à informação** da comunidade baseados em diretrizes internacionais...

Em 23a o substantivo composto **right-to-know** (subst + infinitivo), que significa diretamente, *'right to the knowledge',* aparece em 24b traduzido por **direito à informação,** ou seja, (**$subst_1$ + [prep + art] + $subst_2$).**

24. **dumping-at-sea** (1 ocorrência)

24a. …while maritime transport and **dumping-at-sea** *activities* contribute 10 per cent each.
24b. ... e as *atividades de* transporte marítimo e **descarga no mar** comparecem com 10 por cento cada uma.

Em 24a o nome composto **dumping-at-sea** *(*$subst_1$ + prep + $subst_2$*)* é combinado com a preposição 'at' para indicar '*activities that dump (some kind of waste) at the sea'*, enquanto em

24b o composto foi traduzido por **descarga no mar,** cuja estrutura morfossintática é **(subst$_1$ + [prep + art] + subst$_2$).**

25. **Anti-poverty** (2 ocorrências)

25a. A specific **anti-poverty** *strategy* is therefore one of the basic conditions for ensuring…
25b. Uma estratégia voltada especificamente para o **combate à pobreza**, portanto, é um requisito...

Em 25a, o nome composto **anti-poverty** (pref + subst), *'strategy that combats poverty'*, é traduzido como **combate à pobreza** (25b), cuja estrutura morfossintática é **(subst$_1$ + [prep + art] + subst$_2$).**

26. **post-harvest** (8 ocorrências)

26a …biomass and solar energy to agricultural production and **post-harvest** *activities.*
26b. ...biomassa e à energia solar para a produção agrícola e as atividades **posteriores às colheitas**.

Em 26a, o substantivo composto **post-harvest** (prefixo + subst), que significa '*activities that occur after the harvest'* é traduzido como **posteriores às colheitas,** ou seja**,** (**Adj + [Prep + Art] + Subst).**

SÉTIMO PADRÃO

Nesses casos em que o substantivo composto termina em –ing particípio presente ou –ed particípio passado, o correspondente em português consiste em particípio ligado a um substantivo por uma preposição ou preposição + artigo.

27. **awareness-raising** (4 ocorrências)

27a. Encourage education and **awareness-raising** *programmes* at the local, national, subregional and regional levels concerning energy efficiency…
27b. Fomentar a execução, nos planos local, nacional, sub-regional e regional, de *programas* de ensino e **tomada de consciência** sobre o uso eficiente da energia...

Em 27a, o substantivo composto **awareness-raising** (subst + -ing particípio presente), indica '*programmes that raise awareness'*, enquanto que na tradução (27b) o correspondente é **tomada de consciência**, cuja estrutura morfossintática consiste em (**part + prep + subst**).

28. **community-managed** (1 ocorrência)

28a. Support low-cost, **community-managed** *systems* for the collection of comparable information on the status…
28b. Apoiar *sistemas* de baixo custo, **geridos pela comunidade**, para a coleta de informações comparáveis sobre a situação...

Em 28a, o substantivo composto **community-managed** (subst + -ed particípio passado) significa '*systems that are managed by the community'*, e 28b mostra uma tradução direta para o português, expressa por **geridos pela comunidade***, ou seja,* **(part + prep + subst).**

29. **construction-related** (1 ocorrência)

29a. … against physical disruption by construction and **construction-related** *activities.*
29b. ...dos danos físicos causados pela construção e por atividades **relacionadas à construção**.

Em 29a o substantivo composto **construction-related** (subst + -ed particípio passado) significa '*activities related to construction',* e corresponde a (atividades) **relacionadas à construção (part + [prep + art] + subst).**

30. **forest-related** (5 ocorrências)

30a. …user groups and non-governmental organizations in **forest-related** *activities*, …
30b. ..grupos de usuários e organizações não-governamentais nas atividades **ligadas à floresta**,...

Em 30a o substantivo composto **forest-related** (subst + -ed particípio passado), que significa '*activities related to the forest'*, é traduzido como (atividades) **ligadas à floresta (part + [prep + art] + subst).**

31. **management-related** (60 ocorrências)

31a. These have been internalized into the **management-related** *activities*.
31b. Tudo isso está embutido nas atividades **relacionadas ao manejo**.

Em 31a todas as 60 ocorrências do composto **management-related** (subst + -ed particípio passado), que significa '*activities related to management'*, correspondem a **relacionadas ao manejo (part + [prep + art] + subst).**

32. **water-related** (13 ocorrências)

32a. …of existing institutions in order to enhance their capacities in **water-related** *activities*,…
32b. ...das instituições existentes, com o objetivo de aumentar suas capacidades em atividades **relacionadas com a água**,...

Em 32a o substantivo composto **water-related** (subst + -ed particípio passado), que significa '*activities related to water'*, corresponde em 32b a (atividades) **relacionadas com a água (part + [prep + art] + subst).**

Nesse padrão o segundo elemento é quase sempre *'-related'*, que é traduzido por 'relacionadas' ou 'ligadas'.

OITAVO PADRÃO

Nesse padrão aparece o substantivo composto hifenizado mais frequente no *corpus*, *capacity-building*, cuja tradução correspondente é formada por um substantivo mais um adjetivo, ligado por uma conjunção a outro adjetivo.

33. **capacity-building** (154 ocorrências do *termo em inglês*)

33a. Provide technical and financial assistance for **capacity-building** *programmes* to support the sustainable self-development …
33b. Oferecer assistência técnica e financeira para *programas de* **fortalecimento institucional e técnico** a fim de apoiar o desenvolvimento autônomo sustentável...

Em 33a, o substantivo composto **capacity-building** é formado por substantivo + -ing particípio presente, indicando '*programmes that build capacity*'. Em 33b esse composto foi traduzido pelo sintagma nominal **fortalecimento institucional e técnico,** cuja estrutura morfossintática consiste em (**subst + adj$_1$ + conj + adj$_2$**), ou seja, precisaram ser adicionados dois adjetivos para traduzir o termo do inglês para o português. O uso do termo 'fortalecimento' corresponde a 120 ocorrências.

O termo *capacity-building* consiste em uma tarefa complexa que em inglês é uma unidade institucionalizada. Portanto, foi necessário adicionar dois adjetivos na tradução para o português para especificar o substantivo composto. De acordo com o United Nations Development Program (UNDP) *capacity-building* significa:

> "*… the creation of an enabling environment with appropriate policy and legal frameworks, institutional development, including community participation (of women in particular), human resources development and strengthening of managerial systems, adding that, UNDP recognizes that capacity building is a long-term, continuing process, in which all stakeholders participate (ministries, local authorities, non-governmental organizations and water user groups, professional associations, academics and others (UNDP, 2007).*"

O mesmo termo aparece no *corpus* com uma tradução um pouco diferente. Nesses casos, usou-se o sintagma nominal **capacitação institucional e técnica** (**Subst + Adj_1 + Conj + Adj_2**). Ou seja, também foi necessário adicionar dois adjetivos, mas o substantivo usado foi 'capacitação', em vez de 'fortalecimento'. O termo 'capacitação' foi usando 34 vezes na tradução desse nome composto hifenizado, que foi o único a apresentar traduções diferentes para o português.

NONO PADRÃO

O segundo elemento do substantivo composto hifenizado em inglês consiste em –ed particípio passado, ou substantivo, ou adjetivo, e a estrutura morfossintática da tradução correspondente é mais complexa, sendo sempre iniciada por particípio seguido por preposição, e apresentando em seguida a inserção de tipos pertencentes a várias classes gramaticais (verbo, substantivo, adjetivo, artigo).

34. **forest-related** (5 ocorrências)

34a. …development and implementation of **forest-related** *programmes* and other activities…
34b. ...desenvolvimento e implementação de programas e outras atividades **relacionados à área florestal...**

Em 34a o substantivo composto **forest-related** (subst + -ed particípio passado) significa '*programmes related to forest'*, enquanto que a tradução para o português consiste em **relacionados à área florestal,** cuja estrutura morfossintática é (**part + [prep + art] + subst + adj).**

35. **consumer-oriented** (1 ocorrência)

35a. Encouraging specific **consumer-oriented** *programmes*, such as recycling and deposit…
35b ...programas expressamente **voltados para os interesses do consumidor**, como a reciclagem e sistemas de depósito...

Em 35a o substantivo composto **consumer-oriented** (subst + -ed particípio passado) que indica '*programmes oriented to consumers'* corresponde em português (35b) ao sintagma nominal complexo **voltados para os interesses do consumidor,** formado por (**part + prep$_1$ + art + subst$_1$ + [prep$_2$ + art$_2$] + subst$_2$).**

36. **erosion-control** (1 ocorrência)

36a. Undertake measures to prevent soil erosion and promote **erosion-control** *activities…*
36b. Adotar medidas para evitar a erosão do solo e promover, em todos os setores, atividades **destinadas a controlar a erosão;**

Em 36a o substantivo composto **erosion-control** (subst + subst), que significa '*activities that control erosion'*, corresponde a (atividades) **destinadas a controlar a erosão,** ou seja, **(part + prep** + **verb + art + subst).**

37. **system-wide** (4 ocorrências)

37a. …as well as of **system-wide** *activities* to integrate environment and development,…
37b. assim como das atividades **realizadas em todo o sistema** para integrar meio ambiente e desenvolvimento,…

Em 37a o substantivo composto **system-wide** (subst + adj) corresponde ao complexo sintagma nominal (atividades) **realizadas em todo o sistema (part + prep + adj + art + subst).**

DÉCIMO PADRÃO

Os elementos do composto em inglês incluem sempre um substantivo, e é possível identificar um padrão de tradução que inicia (ou não) por adjetivo, seguido de preposição + pronome + substantivo.

38. **country-specific** (16 ocorrências)

38a. Rather, **country-specific** *programmes* to tackle poverty and international efforts supporting …
38b. ...programas **específicos para cada país**, com atividades internacionais de apoio...

Em 38a, o nome composto **country-specific** (subst + adj), que significa '*programmes specific to (each) country'*, é traduzido em por **específicos para cada país***,* ou seja, (**adj + prep + pron + subst).**

39. **in-country** (2 ocorrências)

39a. Promote **in-country** *programmes* and related physical infrastructure for animal livestock conservation…
39b. Promover, **em seus países**, programas e a infraestrutura física correlata para a conservação...

Em 39a, o substantivo composto **in-country** (prep + subst), que significa '*programmes that occur inside a country'*, corresponde em 39b a **em seus países (prep + pron + subst).**

DÉCIMO PRIMEIRO PADRÃO

Esse padrão inclui casos que não puderam ser encaixados em um dos dez padrões identificados acima. São eles:

40. **early-warning** (6 ocorrências)

40a. …strengthen **early-warning** *systems* and response mechanisms for transboundary air pollution…
40b. ... fortalecer *sistemas de* **pronto alerta** e mecanismos de reação à poluição atmosférica transfronteiriça...

Em 40a, o adjetivo composto **early-warning** (adj + -ing particípio presente) foi usado para indicar '*systems that warn early'*, e sua tradução para o Português foi **pronto alerta**, cuja estrutura morfossintática é **(adj + subst).**

41. **soil/crop-management** (1 ocorrência)

41a….introduce improved **soil/crop-management** *systems* into land-use practice;…
41b. …introduzir melhores sistemas de **manejo terra/cultivo** na prática do uso da terra...

Em 41a, o substantivo composto **soil/crop-management (subst/subst + subst),** que significa '*systems for management of soil/crop'* corresponde em português a **manejo terra/cultivo** (**$subst_1$ + $subst_2$/$subst_3$**).

A seguir, é apresentada uma tabela com os padrões que constituem as regras de tradução para substantivos e adjetivos compostos hifenizados, e a proporção em que eles ocorrem no *corpus*.

Tabela 4 – Informação morfossintática decorrente da tradução de substantivos e adjetivos compostos hifenizados (continua)

PADRÃO (agrupamentos)	*INFORMAÇÃO MORFOSSINTÁTICA (derivada da análise dos quarenta exemplos de nomes compostos hifenizados)*	*%*
1º	subst + -ed → **adj**	6
	adj_1 + adj_2 → **adj**	
2º	$subst_1$ + $subst_2$ → **(prep) + subst**	3
	subst + -ed → **(prep) + subst**	
	subst + -ing → **(prep) + subst**	
3º	adv + -ed → **adv + part**	2
4º	prep + subst → **adv + adj**	2
	adv + subst → **adv + adj**	
5º	subst + -ing → **$subst_1$ + prep + $subst_2$**	10
	subst + -ing → **adj + prep + $subst_2$**	
	adj + -ing → **$subst_1$ + prep + $subst_2$**	
	$subst_1$ + $subst_2$ → **$subst_1$ + prep + $subst_2$**	

Tabela 4 – Informação morfossintática decorrente da tradução de substantivos e adjetivos compostos hifenizados (continuação)

6º	$subst_1$ + $subst_2$ → **$subst_1$ + [prep + art] + $subst_2$**	18
	subst + -ing → **$subst_1$ + [prep + art] + $subst_2$**	
	$subst_1$ + prep + $subst_2$ → **$subst_1$ + [prep + art] + $subst_2$**	
	subst + inf → **$subst_1$ + [prep + art] + $subst_2$**	
	pref + subst → **adj + [prep + art] + $subst_2$**	
7º	subst + -ed → **part + prep + subst**	19
	subst + -ing → **part + prep + subst**	
	subst + -ed → **part + [prep + art] + subst**	
8º	subst + -ing → **subst + adj_1 + conj + adj_2**	33
9º	$subst_1$ + $subst_2$ → **part + prep + verb + art + subst**	2
	subst + adj → **part + prep + adj + art + subst**	
	subst + -ed → **part + [prep + art] + subst + adj**	
	subst + -ed → **part + $prep_1$ + art + $subst_1$ + [$prep_2$ + art_2] + $subst_2$**	
10º	subst + adj → **adj + prep + pron + subst**	3
	prep + subst → **prep + pron + subst**	
11º	adj + -ing → **adj + subst**	2
	$subst_1$/$subst_2$ + $subst_3$ → **$subst_1$ + $subst_2$/$subst_3$**	

Por meio dessa tabela podemos verificar a existência de onze (11) padrões que consistem em vinte e nove (29) formas diferentes de informação morfossintática (considerando o composto hifenizado na língua fonte e sua tradução para a língua alvo), que podem auxiliar a geração de regras de tradução.

Por exemplo, a informação morfossintática resultante da tradução do composto hifenizado *cross-breeding* (subst + -ing → $subst_1$ + prep + $subst_2$) poderia aparecer, de acordo com o formalismo apresentado na Figura 1, da seguinte forma:

SL: Cross-breeding; TL: cruzamento de raças

CN: N ING → N PRE N (informação morfossintática)

Onde *CN* = *compound noun* (informação de tipo)

SL = Source Language; e TL = Target Language

Do mesmo modo, a informação morfossintática resultante da tradução do composto hifenizado *construction-related* (subst + -ed → part + [prep + art] + subst) apareceria como:

SL: construction-related; TL: relacionadas à construção
CN: N ED → PART PRE ART N (Informação morfossintática)

Desse modo, a estrutura morfossintática relativa a cada um dos exemplos apresentados na Tabela 4 pode ser utilizada para compor o formalismo necessário para a geração de regras de tradução, em sistemas de tradução automática.

4.3 Resultados da análise das características do *corpus* de estudo

Os resultados da contagem das ocorrências do *corpus* são mostrados na Tabela 5 a seguir, que apresenta a frequência (valores) e percentagem das ocorrências para cada classe gramatical, em ordem decrescente.

Tabela 5 – Distribuição das ocorrências por classe gramatical (continua)

Classe gramatical	***Ocorrências (valores)***	***Ocorrências (percentagem - %)***
SUBSTANTIVO	49.818	34
PREPOSIÇÃO	24.328	17
ADJETIVO	23.162	16
VERBO	12.716	9
CONJUNÇÃO	12.460	8
ARTIGO	10.083	7
PRONOME	4.669	3
ADVÉRBIO	3.846	3
VERBO AUXILIAR	3.106	2

Tabela 5 – Distribuição das ocorrências por classe gramatical (continuação)

MODAL	2.184	1
ACRÔNIMO	431	0
PREFIXO	250	0
OUTROS IDIOMAS	222	0
ERRO	83	0
ELEMENTO DE COMPOSIÇÃO	44	0
TOTAL	**147.365**	**100**

Por meio da tabela acima e da Figura 2 pode-se constatar que as 3 frequências mais altas correspodem às classes de substantivo, preposição e adjetivo (49.818, 24.328 e 23.162, respectivamente). A frequência das conjunções – palavras invariáveis que ligam duas orações ou dois termos semelhantes da mesma oração – e verbos é muito semelhante (12.460 e 12.716). A seguir, em ordem decrescente de frequência temos os artigos, pronomes, advérbios, verbos auxiliares, modais, acrônimos, prefixos, palavras pertencentes a outros idiomas, erros (que incluem possíveis erros de limpeza do *corpus* e de contagem das ocorrências), e por fim elementos de composição. Nos dados acima não está incluído o aparecimento de uma sobrecontagem (provavelmente devido à dupla contagem da ocorrência de um mesmo tipo) que perfaz 37 ocorrências, ou seja, a soma total das classes foi de 147.402 ocorrências). Considera-se essa diferença admissível, visto o tamanho do *corpus* de estudo (147.635).

O aparecimento de erros, visto que perfazem um número muito pequeno em relação às principais classes gramaticais, não invalida a análise e não afeta o objetivo principal da abordagem baseada em *corpus*: confirmar ou rejeitar as características de documentos oficiais, apresentadas por Biber (1988).

Pode-se observar através da leitura do *corpus* um caráter descritivo – que consiste na descrição da problemática ambiental – e altamente prescritivo (conforme mencionado anteriormente), ao determinar as ações a serem tomadas para minimizar esses problemas. Alguns exemplos de tais características do discurso do *corpus* são mostrados mais adiante.

Frequência das ocorrências por classe gramatical

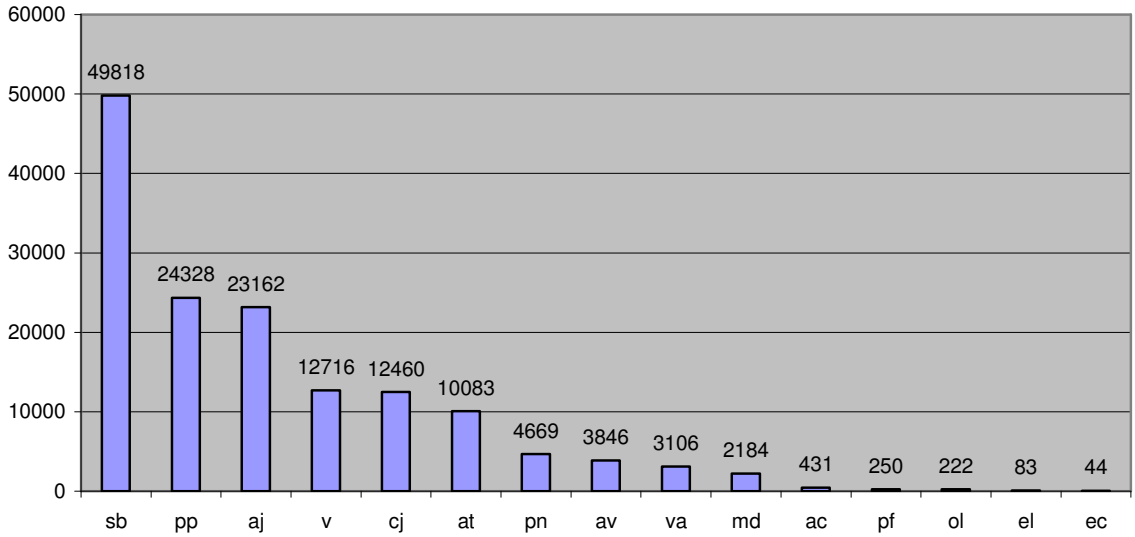

Figura 2 – Distribuição das ocorrências por classe gramatical (valores).

O número de conjunções pode ser explicado devido à ocorrência de frases muito longas, que precisam ser ligadas através desses itens gramaticais.

A Figura 3 abaixo mostra a percentagem das ocorrências por classe gramatical. Observa-se que os itens gramaticais perfazem 35%, enquanto que a percentagem de itens lexicais soma 65%.

As preposições e substantivos, juntos, apresentam uma frequência de 51%, ou seja, praticamente a metade de todas as ocorrências computadas. A percentagem de adjetivos é de 16%, conjunções e verbos de 9%, e artigos 7%.

As demais percentagens consistem em: pronomes (3%), advérbios (3%), verbos auxiliares (2%), modais (1%). As demais percentagens são praticamente iguais 0%.

A partir das percentagens mostradas na Figura 3 – que mostram as proporções relativas às classes gramaticais existentes no *corpus* – e da leitura do *corpus* foi possível confirmar a maior parte das características de documentos oficiais apresentadas por Biber (1988), e ainda detectar outras características específicas do *corpus* de estudo.

Percentagem das ocorrências por classe gramatical

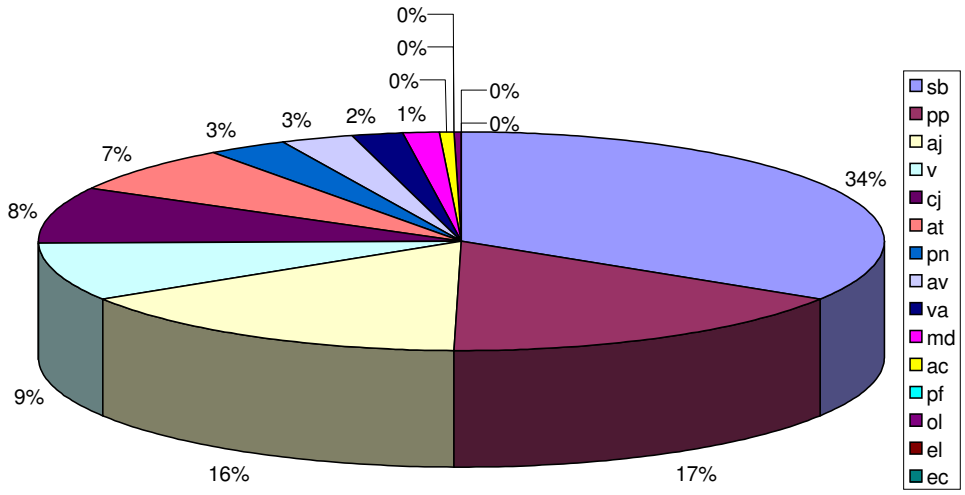

Figura 3 – Percentagem das ocorrências por classe gramatical

Tais características do *corpus* de estudo são apresentadas abaixo:

- alta frequência de preposições e substantivos, considerando todo o *corpus* de estudo; tais palavras perfazem 51% das ocorrências, a metade de todas as ocorrências computadas. Essa constatação está de acordo com Biber (1988).

- *vários* adjetivos atributivos (em inglês, adjetivo que precede o substantivo, atribuindo-lhe qualidade, caráter ou estado); perfazendo 16% das ocorrências. Essa também é uma característica apontada por Biber (1988) e confirmada pelo exame do *corpus* de estudo. Os trechos abaixo mostram o uso de vários adjetivos atributivos numa mesma sentença:

> *"A **specific anti-poverty strategy** is therefore one of the **basic conditions** for ensuring **sustainable development.**"* Ou em:

*"Poverty is a **complex multidimensional problem** with origins in both the **national and international domains**."*

- frases muito longas também podem ser observadas, conforme mostrado abaixo:

"Examine the international economic framework, including resource flows and structural adjustment programmes, to ensure that social and environmental concerns are addressed, and in this connection, conduct a review of the policies of international organizations, bodies and agencies, including financial institutions, to ensure the continued provision of basic services to the poor and needy."

- muitos *verbos* são usados no infinitivo, mostrando os objetivos e ações recomendadas pelo documento; por exemplo em:

*"**To create** a focus in national development plans and budgets on investment in human capital,…"*
*"**To provide** all persons urgently with the opportunity **to earn** a sustainable livelihood…"*

- o discurso é descritivo, ao fazer um relato de fatos relacionados à problemática ambiental; e também prescritivo, ao determinar o que deve ser feito para resolver os problemas relacionados ao meio ambiente. Isso pode ser constatado nos trechos mostrados abaixo:

"Health ultimately depends on the ability to manage successfully the interaction between the physical, spiritual, biological and economic/social environment. Sound development is not possible without a healthy population; yet most developmental activities affect the environment to some degree, which in turn causes or exacerbates many health problems." (descritivo)

*"Scientific and technological means new approaches to planning and managing health care systems and facilities **should** be tested, and research on ways of*

integrating appropriate technologies into health infrastructures supported. The development of scientifically sound health technology ***should*** *enhance adaptability to local needs and maintainability by community resources, including the maintenance and repair of equipment used in health care." (prescritivo)*

- uso de muitos verbos modais, vinculados ao caráter prescritivo do documento – além de *should,* como também aparece acima, ocorrem também os modais *must, e will* – precedendo verbos. Essa é uma característica não observada por Biber em documentos oficiais.

  "These approaches ***should*** *form the core principles of national settlement strategies. In developing these strategies, countries* ***will*** *need to set priorities among the eight programme areas in this document in accordance with their national plans and objectives taking fully into account their social and cultural capabililities."*

  "Increasing the efficiency of energy use to reduce its polluting effects and to promote the use of renewable energies ***must*** *be a priority in any action taken to protect the urban environment."*

- presença de 8% de conjunções, uma percentagem significativa, que pode estar relacionada em parte ao uso de sentenças muito longas. Tais sentenças foram largamente observadas no documento analisado. O trecho abaixo mostra o uso da conjunção *"and"* quatro vezes na mesma sentença.

  "The objectives are to extend the provision of more energy-efficient technology ***and*** *alternative/renewable energy for human settlements* ***and*** *to reduce negative impacts of energy production* ***and*** *use on human health* ***and*** *on the environment."*

- não foram encontradas muitas construções passivas no documento oficial analisado, contrariando Biber (1988).

5 Conclusões

De acordo com os resultados apresentados, podem ser tiradas algumas conclusões sobre as características do *corpus* de estudo e acerca da tradução dos substantivos e adjetivos compostos hifenizados.

A partir de exemplos extraídos de um *corpus* paralelo, foram analisadas as correspondências de tradução para quarenta e um exemplos de substantivos e adjetivos compostos hifenizados. Tais exemplos foram agrupados dentro de onze padrões de tradução, que resultaram em vinte e nove estruturas morfossintáticas diferentes, considerando o composto hifenizado na língua fonte e sua tradução para a língua alvo.

Observou-se que no processo de tradução inglês-português dos nomes compostos hifenizados ocorreu a omissão de certos tipos, como substantivos ou elementos terminados em –ed ou –ing. Foi observada a reordenação de palavras na tradução dos nomes compostos para o português (na sentença alvo), com ou sem mudança de classe gramatical, juntamente com a inserção de preposições acompanhadas ou não por artigos. Em alguns exemplos também foram inseridos outros tipos como pronomes, conjunções, verbos, além de substantivos e/ou adjetivos adicionais. Em vista do tipo de processamento linguístico necessário para a tradução de nomes compostos hifenizados, é possível afirmar que a informação morfossintática resultante da análise feita por esta pesquisa pode auxiliar a geração de regras de tradução em sistemas baseados em exemplos (EBMT). Tal informação pode ser aproveitada em sistemas de tradução automática direta, especialmente em *corpora* de domínio restrito; ou em sistemas híbridos, de modo que cada método seja responsável por determinados aspectos do processo de tradução.

Quanto às características do *corpus*, pode-se constatar que as preposições e substantivos perfazem mais da metade das ocorrências do *corpus* (51%).

Através da percentagem de adjetivos (frequência de 16%) e da leitura do *corpus* constatou-se que tais adjetivos eram usados atributivamente. Adicionalmente, uma das características observadas – a presença de 8% de conjunções – pode estar relacionada ao uso de sentenças muito longas, as quais foram largamente utilizadas no documento analisado.

Dado o caráter altamente prescritivo do documento, destaca-se também uma alta freqüência de verbos modais. Por sua vez, os verbos usados no infinitivo ocorrem em grande

parte no início das frases que determinam os objetivos de cada capítulo do documento e recomendam ações a serem tomadas. Foram observadas poucas construções passivas.

Já o caráter descritivo do *corpus* de estudo transparece no relato extensivo sobre os problemas relacionados ao meio ambiente.

Portanto, a análise das características do *corpus* de estudo confirma a maior parte das características mencionadas por Biber, e ainda fornece outras características específicas ao documento oficial em questão. Tais características referem-se ao aparecimento de muitos verbos modais, associados ao caráter prescritivo do documento oficial, e ao uso significativo de conjunções, que pode revelar uma função de ligação das longas sentenças que caracterizam o *corpus* de estudo.

Referências

AIJMER, Karin; ALTENBERG, Bengt;. The english-swedish parallel corpus: a resource for contrastive research and translation studies. In: MAIR, Christian.; HUNDT, Marianne. **Corpus Linguistics and linguistic theory**. Rodopi, 2000, p. 15-17.

HUNDT, Marianne. **Corpus Linguistics and linguistic theory**. Rodopi, 2000, p. 15-17.

BAKER, Mona. "*Linguistics and Cultural Studies. Complementary or Competing Paradigms in Translation Studies*?".A. Lauer et al. (orgs.) In: *Ubersetzungswissenschaft im Umbruch*, Tubingen: Gunter Narr, 1996.

BAKER, Mona (ed.) **Encyclopedia of Translation Studies**. London: Routledge, 1998, 680p.

BERBER SARDINHA, T. B. ***Corpora eletrônicos na pesquisa em tradução***. In: Tagnin, S. E. O. (Org.). Cadernos de Tradução: Corpora e Tradução. Florianópolis: NUT, 2002, v. 1, n. 9, p. 15-59.

_____ ***O uso de corpora na formação de tradutores***. **DELTA**. São Paulo, v. 29, n. spe, pp. 43-70, 2003.

_____**Lingüística de Corpus**. São Paulo: Manole, 2004, 410 p.

_____**Como encontrar as palavras-chave mais importantes de um corpus com WordSmith tools.** DELTA, v. 21, n. 2, 2005, p. 237-250.

BIBER, Douglas. **Variation across speech** and **writing**. Cambridge: Cambridge University Press, 1988. 299p.

BOSTRÖM, H. (2000) Induction of recursive transfer rules. IN: CUSSENS, J.; DEROSKI, S.; (eds.), **Learning Language in Logic, Lecture Notes in Computer Science**, v. 1925, Springer-Verlag Heidelberg, p. 237-246.

BOWKER, L. 2000. Towards a Methodology for Exploiting Specialized Target Language Corpora as Translation Resources. **International Journal of Corpus Linguistics,** Vol. 5(1), 17-52.

CARBONNEL, J.; PROBST, K.; PETERSON, E.; MONSON, C. ; LAVIE, A. ; BROWN, R. LEVIN. L. (2002). Automatic rule learning for resource-limited MT. IN: *Proceedings of the Fifth Conference of the Association for machine Translation in the Americas* (AMTA 2002). Tiburon, California, p. 1-10.

CARTER, R.; MCCARTHY, M. Cambridge Grammar of English, 2006.

CASELI, Helena de Medeiros. **Regras de tradução automáticas induzidas de textos paralelos envolvendo o português do Brasil.** (Projeto de Qualificação Doutorado) Instituto de Ciências matemáticas e Computação, USP, 2004, 67 p.

DORR, B.J.; JORDAN, P.W.; BENOIT, J.W. **A survey of current paradigms in machine translation.** IN: M. Zelkowitz (ed.) Advances in computers, v. 49, London: Academic press, pp. 1-68.

FLOWERDEW, Lynne. **The argument for using English specialized corpora to understand academic and professional language**. IN: CONNOR, Ulla; UPTON, Thomas Albin. Discourse In The Professions: Perspectives From Corpus Linguistics, John Benjamins, 2004, p. 11-36.

CASELI, Helena de Medeiros. **Regras de tradução automáticas induzidas de textos paralelos envolvendo o português do Brasil.** (Projeto de Qualificação Doutorado) Instituto de Ciências matemáticas e Computação, USP, 2004, 67 p.

CARBONNEL, J.; PROBST, K.; PETERSON, E.; MONSON, C. ; LAVIE, A. ; BROWN, R. LEVIN. L. (2002). Automatic rule learning for resource-limited MT. IN: *Proceedings of the Fifth Conference of the Association for machine Translation in the Americas* (AMTA 2002). Tiburon, California, p. 1-10.

GRANGER, Sylviane. **The corpus approach: a common way forward for Contrastive Linguistics and Translation Studies?** IN: GRANGER; LEROT; PETCH-TYSON. Corpus-based approach to contrastive linguistics and translation Studies. Rodopi, 2003, p. 17-30.

HALLIDAY, M. A. K. Writing science: literacy and discursive power**. Pittsburg: Routledge, 1993.**

HOEY, M. Patterns of lexis in text. Oxford University Press, 1991.

HUNSTON, S. Corpora and applied linguistics. Cambridge: Cambridge University Press, 2002.

HUTCHINS, John. Machine translation. General Overview. IN: Mitkov, Ruslan (ed.) **The Oxford Handbook of Computational Linguistics**. Oxford: University Press, 2003, p. 501-511.

JOHANSSON, Stig; OKSEFJELL, Signe. **Corpora and Cross-linguistic Research: Theory, Method and Case Studies,** RODOPI, 1998, 376p.

LAVIOSA, S. The English Comparable Corpus: **A Resource and a Methodology.** IN: Bowker L., Cronin M., Kenny D. and J. Pearson (ed.) *Unity in Diversity: Current Trends in Translation Studies*. Manchester: St. Jerome Publishing, 1998.

LAVIOSA, Sara. **Corpus-based Translation Studies.** Theory, findings and applications. RODOPI, 2002, 148p.

MATSUMOTO, Y.; ISHIMOTO, H.; UTSURO, T. (1993). Structural matchinf of parallel texts. *IN: Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics* (ACL'93), Columbus, Ohio. P. 23-30.

_____ Corpora and Translation Studies. IN: GRANGER; LEROT; PETCH-TYSON. **Corpus-based approach to contrastive linguistics and translation Studies.** Rodopi, 2003, p. 45-56.

LEECH, Geoffrey. **Semantics, Penguins, Harmondsworth**, 1974.

LEECH, Geoffrey. Corpora and theories of linguistic performance. IN: **Directions in Corpus Linguistics.** Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991.

McENERY, T.; WILSON, A. **Corpus Linguistics***,* Edinburgh University Press, 1996.

_____ **Corpus Linguistics**, Edinburgh 2nd edition, Edinburgh University Press, 2001

PARTINGTON, A. **Patterns and Meanings**, Amsterdan: John Benjamins, 1998.

SCOTT, M. Comparing corpora and identifying key words, collocations, and frequency distributions. **IN: GHADESSY, M.; HENRY, A.; ROSEBERRY, R. L.** Small Corpus Studies and Elt: Theory and Practice**, Amsterdan: John Benjamins, 2001.**

RAYSON, P.; GARSIDE, R. (2000). **Comparing corpora using frequency profiling**. In: Proceedings of the workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000). 1-8 October 2000, Hong Kong, pp. 1 - 6

RAYSON P.; BERRIDGE D.; FRANCIS B. (2004). **Extending the Cochran rule for the comparison of word frequencies between corpora.** In:*Volume II of Purnelle G., Fairon C., Dister A. (eds.) Le poids des mots: Proceedings of the 7th International Conference on Statistical analysis of textual data (JADT 2004), Louvain-la-Neuve, Belgium, March 10-12, 2004*, Presses universitaires de Louvain, pp. 926 - 936.

SINCLAIR, J. **Corpus, concordance and collocations**. Oxford University Press, 1991.

TAGNIN, S. E. O. Os **Corpora: instrumentos de auto-ajuda para o tradutor**. In: Tagnin, S. E. O. (Org.). Cadernos de Tradução: Corpora e Tradução. Florianópolis: NUT, 2002, v. 1, n. 9, p. 191-218.

THUNES, M.. Classifying translational correspondences. IN: JOHANSSON, S.; OKSEFJELL. **Corpora and Cross-linguistic Research**. Theory, Method and Case Studies. Amsterdam: RODOPI, 1998. p. 3-51.

TOGNINI-BONELLI, E. **Corpus linguistics at work**. Amsterdan: John Benjamins, 2001, 223pp.

ANEXO A

Lista de palavras-chave

N	WORD	FREQ.	AG21.LST %	FREQ.	RITT.LST %	KEYNESS	P
1	AND	*11.187*	*7,27*	32.815	2,68	8.508,6	0,000000
2	DEVELOPMENT	*1.403*	*0,91*	31.114	0,03	6.458,0	0,000000
3	SUSTAINABLE	*613*	*0,40*	654		6.072,6	0,000000
4	ORGANIZATIONS	*736*	*0,48*	2.799		5.789,0	0,000000
5	INTERNATIONAL	*1.000*	*0,65*	21.761	0,02	4.636,2	0,000000
6	APPROPRIATE	*818*	*0,53*	10.863	0,01	4.553,0	0,000000
7	PROGRAMMES	*704*	*0,46*	6.212		4.457,8	0,000000
8	GOVERNMENTS	*616*	*0,40*	4.570		4.098,7	0,000000
9	COUNTRIES	*826*	*0,54*	16.230	0,02	3.987,5	0,000000
10	RESOURCES	*681*	*0,44*	9.930	0,01	3.669,9	0,000000
11	ENVIRONMENTAL	*633*	*0,41*	7.968		3.586,4	0,000000
12	ACTIVITIES	*692*	*0,45*	11.316	0,01	3.579,6	0,000000
13	SHOULD	*1.401*	*0,91*	96.455	0,11	3.536,4	0,000000
14	COOPERATION	*416*	*0,27*	1.248		3.442,8	0,000000
15	REGIONAL	*597*	*0,39*	7.425		3.395,8	0,000000
16	IMPLEMENTATION	*450*	*0,29*	2.799		3.139,9	0,000000
17	DEVELOPING	*527*	*0,34*	5.972		3.089,6	0,000000
18	MANAGEMENT	*712*	*0,46*	20.934	0,02	2.897,9	0,000000
19	INCLUDING	*707*	*0,46*	23.694	0,03	2.705,1	0,000000
20	NATIONAL	*782*	*0,51*	36.493	0,04	2.515,8	0,000000
21	CONCESSIONAL	*203*	*0,13*	25		2.433,3	0,000000
22	NATIONS	*397*	*0,26*	4.028		2.409,5	0,000000
23	B	*640*	*0,42*	22.383	0,02	2.397,9	0,000000
24	ENVIRONMENTALL+	*264*	*0,17*	665		2.262,9	0,000000
25	ENVIRONMENT	*509*	*0,33*	12.417	0,01	2.249,0	0,000000
26	PROMOTE	*343*	*0,22*	3.047		2.166,8	0,000000
27	RELEVANT	*407*	*0,26*	7.677		1.995,1	0,000000
28	WASTES	*227*	*0,15*	617		1.916,6	0,000000
29	GOVERNMENTAL	*231*	*0,15*	900		1.806,3	0,000000
30	RESOURCE	*278*	*0,18*	2.578		1.733,8	0,000000
31	DEVELOP	*374*	*0,24*	8.274		1.720,9	0,000000
32	TECHNOLOGIES	*245*	*0,16*	1.523		1.709,4	0,000000
33	PROGRAMME	*465*	*0,30*	17.703	0,02	1.668,7	0,000000
34	CAPACITY	*325*	*0,21*	5.721		1.635,3	0,000000
35	C	*511*	*0,33*	25.736	0,03	1.572,4	0,000000
36	SUPPORT	*534*	*0,35*	28.950	0,03	1.572,3	0,000000
37	STRENGTHEN	*216*	*0,14*	1.168		1.562,0	0,000000
38	ALIA	*175*	*0,11*	381		1.542,2	0,000000
39	STRATEGIES	*246*	*0,16*	2.713		1.454,3	0,000000
40	OBJECTIVES	*261*	*0,17*	4.109		1.368,1	0,000000
41	INFORMATION	*540*	*0,35*	36.912	0,04	1.367,5	0,000000
42	STRENGTHENING	*177*	*0,11*	762		1.352,6	0,000000
43	SCIENTIFIC	*284*	*0,18*	5.919		1.338,5	0,000000
44	WATER	*503*	*0,33*	33.631	0,04	1.293,2	0,000000
45	UNITED	*390*	*0,25*	18.210	0,02	1.253,4	0,000000
46	ECONOMIC	*431*	*0,28*	23.883	0,03	1.251,8	0,000000
47	MECHANISMS	*203*	*0,13*	1.967		1.249,3	0,000000
48	MARINE	*207*	*0,13*	2.145		1.247,9	0,000000
49	HAZARDOUS	*161*	*0,10*	721		1.219,0	0,000000
50	D	*339*	*0,22*	13.294	0,01	1.197,8	0,000000

N	WORD	FREQ.	AG21.LST %	FREQ.	RITT.LST %	KEYNESS	P
51	COORDINATION	*145*	*0,09*	445		1.194,1	0,000000
52	POLICIES	*288*	*0,19*	8.250		1.185,5	0,000000
53	NON	*401*	*0,26*	21.704	0,02	1.181,6	0,000000
54	SUBREGIONAL	*97*	*0,06*	11		1.166,9	0,000000
55	GLOBAL	*221*	*0,14*	3.546		1.150,5	0,000000
56	WASTE	*260*	*0,17*	6.386		1.145,1	0,000000
57	INDIGENOUS	*162*	*0,11*	971		1.141,1	0,000000
58	HUMAN	*371*	*0,24*	19.380	0,02	1.116,9	0,000000
59	CAPACITIES	*133*	*0,09*	405		1.097,1	0,000000
60	INTEGRATED	*197*	*0,13*	2.632		1.093,5	0,000000
61	LAND	*364*	*0,24*	20.231	0,02	1.055,1	0,000000
62	IMPLEMENTING	*147*	*0,10*	832		1.050,8	0,000000
63	RELATED	*279*	*0,18*	9.749	0,01	1.045,2	0,000000
64	AGENDA	*175*	*0,11*	1.948		1.031,4	0,000000
65	TECHNOLOGICAL	*166*	*0,11*	1.644		1.014,8	0,000000
66	CONSERVATION	*204*	*0,13*	3.873		997,2	0,000000
67	SYSTEMS	*329*	*0,21*	17.011	0,02	996,4	0,000000
68	INTER	*188*	*0,12*	2.901		992,5	0,000000
69	SECRETARIAT	*123*	*0,08*	422		989,2	0,000000
70	AREAS	*358*	*0,23*	21.953	0,02	974,9	0,000000
71	ECOSYSTEMS	*101*	*0,07*	130		972,9	0,000000
72	UNEP	*93*	*0,06*	106		912,3	0,000000
73	MEASURES	*224*	*0,15*	6.774		899,3	0,000000
74	OF	*7.195*	*4,67*	37.887	3,24	892,8	0,000000
75	USE	*519*	*0,34*	57.805	0,06	885,9	0,000000
76	RESEARCH	*362*	*0,24*	26.533	0,03	872,3	0,000000
77	PROMOTING	*143*	*0,09*	1.499		858,9	0,000000
78	HEALTH	*340*	*0,22*	23.627	0,03	850,8	0,000000
79	LEVELS	*259*	*0,17*	11.851	0,01	841,8	0,000000
80	ESTABLISH	*198*	*0,13*	5.311		839,3	0,000000
81	SOUND	*274*	*0,18*	13.951	0,02	837,4	0,000000
82	TRAINING	*308*	*0,20*	19.101	0,02	832,5	0,000000
83	ASSESSMENT	*210*	*0,14*	6.602		827,9	0,000000
84	INSTITUTIONS	*208*	*0,14*	6.421		827,0	0,000000
85	IMPLEMENT	*138*	*0,09*	1.469		824,9	0,000000
86	UTILIZATION	*84*	*0,05*	101		817,5	0,000000
87	COMMUNITIES	*178*	*0,12*	4.026		811,4	0,000000
88	DESERTIFICATIO+	*79*	*0,05*	83		784,0	0,000000
89	TECHNICAL	*202*	*0,13*	6.630		780,1	0,000000
90	LEVEL	*340*	*0,22*	27.093	0,03	770,1	0,000000
91	EXISTING	*223*	*0,14*	9.442	0,01	756,4	0,000000
92	EVALUATION	*153*	*0,10*	2.891		749,2	0,000000
93	COASTAL	*127*	*0,08*	1.428		746,1	0,000000
94	INDICATIVE	*99*	*0,06*	509		724,9	0,000000
95	SPECIFIC	*230*	*0,15*	11.577	0,01	707,6	0,000000
96	IMPACTS	*85*	*0,06*	252		705,0	0,000000
97	DATA	*271*	*0,18*	18.051	0,02	698,2	0,000000
98	PARTICIPATION	*141*	*0,09*	2.681		688,8	0,000000
99	LOCAL	*395*	*0,26*	43.330	0,05	683,7	0,000000
100	DEGRADATION	*92*	*0,06*	466		676,1	0,000000

N	WORD	FREQ.	AG21.LST %	FREQ.	RITT.LST %	KEYNESS	P
101	BIOLOGICAL	*126*	*0,08*	1.952		664,2	0,000000
102	COST	*269*	*0,17*	20.163	0,02	637,6	0,000000
103	MAGNITUDE	*102*	*0,07*	972		630,8	0,000000
104	PLANNING	*225*	*0,15*	13.531	0,01	620,4	0,000000
105	FINANCIAL	*240*	*0,16*	15.911	0,02	620,3	0,000000
106	ENSURE	*198*	*0,13*	9.930	0,01	610,5	0,000000
107	ACTION	*269*	*0,17*	21.447	0,02	608,9	0,000000
108	CHEMICALS	*121*	*0,08*	2.129		608,8	0,000000
109	COMMUNITY	*269*	*0,17*	21.748	0,02	602,5	0,000000
110	MULTILATERAL	*77*	*0,05*	302		601,1	0,000000
111	FINANCING	*101*	*0,07*	1.194		583,8	0,000000
112	UPON	*268*	*0,17*	22.820	0,03	576,6	0,000000
113	BASIS	*215*	*0,14*	13.511	0,01	576,0	0,000000
114	ERGOVERNMEN+	*70*	*0,05*	222		572,5	0,000000
115	MEANS	*275*	*0,18*	24.478	0,03	571,0	0,000000
116	FACILITATE	*95*	*0,06*	998		570,2	0,000000
117	RURAL	*159*	*0,10*	6.109		567,6	0,000000
118	POLLUTION	*138*	*0,09*	4.097		558,7	0,000000
119	INSTITUTIONAL	*111*	*0,07*	2.005		553,0	0,000000
120	REVIEWED	*100*	*0,06*	1.436		541,4	0,000000
121	EFFECTIVE	*183*	*0,12*	9.963	0,01	537,1	0,000000
122	NEEDS	*234*	*0,15*	18.432	0,02	534,5	0,000000
123	BIOTECHNOLOGY	*71*	*0,05*	348		526,0	0,000000
124	TERMS	*263*	*0,17*	24.570	0,03	524,7	0,000000
125	URBAN	*145*	*0,09*	5.457		523,2	0,000000
126	TECHNOLOGY	*191*	*0,12*	11.657	0,01	521,6	0,000000
127	IMPROVE	*147*	*0,10*	5.825		516,4	0,000000
128	DROUGHT	*80*	*0,05*	673		513,4	0,000000
129	REUSE	*51*	*0,03*	53		506,9	0,000000
130	TRANSBOUNDARY	*45*	*0,03*	15		506,9	0,000000
131	ESTIMATES	*109*	*0,07*	2.402		502,2	0,000000
132	BUILDING	*224*	*0,15*	18.195	0,02	499,9	0,000000
133	ENCOURAGE	*131*	*0,09*	4.840		477,1	0,000000
134	PRODUCTION	*202*	*0,13*	15.430	0,02	472,1	0,000000
135	AGENCIES	*117*	*0,08*	3.543		469,3	0,000000
136	DEPEND	*113*	*0,07*	3.332		458,9	0,000000
137	DISPOSAL	*98*	*0,06*	2.111		455,7	0,000000
138	AGRICULTURAL	*117*	*0,08*	3.828		452,5	0,000000
139	MONITORING	*104*	*0,07*	2.709		446,5	0,000000
140	INSTRUMENTS	*105*	*0,07*	2.823		444,5	0,000000
141	SANITATION	*53*	*0,03*	154		441,4	0,000000
142	ENHANCE	*85*	*0,06*	1.385		440,0	0,000000
143	FORESTS	*93*	*0,06*	1.945		437,6	0,000000
144	CONFERENCE	*157*	*0,10*	9.375	0,01	434,8	0,000000
145	EFFORTS	*128*	*0,08*	5.465		432,1	0,000000
146	IMPROVING	*96*	*0,06*	2.484		413,3	0,000000
147	PRACTICES	*115*	*0,07*	4.463		408,4	0,000000
148	ASSISTANCE	*112*	*0,07*	4.218		403,9	0,000000
149	COOPERATE	*55*	*0,04*	303		395,8	0,000000
150	ADEQUATE	*103*	*0,07*	3.439		394,5	0,000000

N	WORD	FREQ.	AG21.LST %	FREQ.	RITT.LST %	KEYNESS	P
151	TRANSFER	*130*	*0,08*	6.730		393,3	0,000000
152	ACCESS	*154*	*0,10*	10.491	0,01	390,6	0,000000
153	PARTICULARLY	*206*	*0,13*	20.465	0,02	389,8	0,000000
154	ESTIMATED	*113*	*0,07*	4.870		379,5	0,000000
155	CAPABILITIES	*68*	*0,04*	896		379,2	0,000000
156	AWARENESS	*102*	*0,07*	3.621		379,0	0,000000
157	COLLABORATION	*75*	*0,05*	1.326		376,6	0,000000
158	PROTECTION	*133*	*0,09*	7.789		372,9	0,000000
159	INFRASTRUCTURE	*68*	*0,04*	942		372,8	0,000000
160	DEMOGRAPHIC	*63*	*0,04*	744		364,3	0,000000
161	SECTOR	*136*	*0,09*	8.569		363,7	0,000000
162	ACCOUNT	*176*	*0,11*	15.876	0,02	361,4	0,000000
163	ENERGY	*157*	*0,10*	12.281	0,01	360,5	0,000000
164	SECTORAL	*43*	*0,03*	126		357,5	0,000000
165	STATES	*183*	*0,12*	17.604	0,02	356,1	0,000000
166	F	*128*	*0,08*	7.690		353,1	0,000000
167	E	*198*	*0,13*	21.235	0,02	349,8	0,000000
168	SOCIO	*67*	*0,04*	1.070		349,3	0,000000
169	ORDER	*248*	*0,16*	33.370	0,04	349,1	0,000000
170	TOXIC	*70*	*0,05*	1.279		347,2	0,000000
171	NETWORKS	*78*	*0,05*	1.883		346,0	0,000000
172	COSTS	*164*	*0,11*	14.331	0,02	345,5	0,000000
173	MINIMIZATION	*32*	*0,02*	18		343,1	0,000000
174	DIVERSITY	*70*	*0,05*	1.380		337,2	0,000000
175	PROVIDE	*195*	*0,13*	21.430	0,02	336,8	0,000000
176	AGRICULTURE	*95*	*0,06*	3.722		335,7	0,000000
177	SUSTAINABILITY	*41*	*0,03*	133		333,7	0,000000
178	SECTORS	*79*	*0,05*	2.160		332,0	0,000000
179	BILATERAL	*59*	*0,04*	769		330,2	0,000000
180	PREVENTION	*70*	*0,05*	1.460		329,7	0,000000
181	REVIEW	*131*	*0,09*	9.142	0,01	326,7	0,000000
182	FOR	*2.142*	*1,39*	833.332	0,92	326,4	0,000000
183	METHODOLOGIES	*39*	*0,03*	114		324,4	0,000000
184	GUIDELINES	*78*	*0,05*	2.186		324,2	0,000000
185	RECYCLING	*60*	*0,04*	918		317,7	0,000000
186	DISSEMINATION	*47*	*0,03*	336		315,8	0,000000
187	FOREST	*114*	*0,07*	6.958		311,2	0,000000
188	POPULATIONS	*67*	*0,04*	1.461		310,0	0,000000
189	PROCESSES	*102*	*0,07*	5.363		305,7	0,000000
190	ECOLOGICAL	*55*	*0,04*	732		305,6	0,000000
191	UNDERTAKE	*69*	*0,04*	1.679		305,0	0,000000
192	PRIORITIES	*71*	*0,05*	1.862		303,9	0,000000
193	ACTUAL	*104*	*0,07*	5.704		303,8	0,000000
194	UNCED	*35*	*0,02*	89		299,4	0,000000
195	SETTLEMENTS	*61*	*0,04*	1.149		299,0	0,000000
196	BIODIVERSITY	*35*	*0,02*	95		295,5	0,000000
197	PGRFA	*23*	*0,01*	0		293,5	0,000000
198	FRESHWATER	*44*	*0,03*	323		293,5	0,000000
199	SOURCES	*107*	*0,07*	6.545		291,7	0,000000
200	EXCHANGE	*120*	*0,08*	8.723		290,7	0,000000

N	WORD	FREQ.	AG21.LST %	FREQ.	RITT.LST %	KEYNESS	P
201	EDUCATION	*194*	*0,13*	24.967	0,03	286,4	0,000000
202	IMPACT	*109*	*0,07*	7.138		284,0	0,000000
203	IMPROVED	*92*	*0,06*	4.662		281,9	0,000000
204	DECIDE	*101*	*0,07*	5.975		281,4	0,000000
205	FRAMEWORK	*85*	*0,06*	3.816		279,1	0,000000
206	PROCEDURES	*93*	*0,06*	4.945		276,9	0,000000
207	INTEGRATE	*49*	*0,03*	649		272,7	0,000000
208	DEVELOPMENTAL	*48*	*0,03*	620		269,4	0,000000
209	EFFICIENT	*84*	*0,05*	3.977		267,7	0,000000
210	BILLION	*88*	*0,06*	4.743		259,8	0,000000
211	NECESSARY	*154*	*0,10*	17.392	0,02	259,2	0,000000
212	FISHERIES	*49*	*0,03*	761		258,0	0,000000
213	CONSUMPTION	*76*	*0,05*	3.244		256,6	0,000000
214	INCENTIVES	*52*	*0,03*	991		253,8	0,000000
215	INTEGRATION	*68*	*0,04*	2.393		253,8	0,000000
216	GRANT	*102*	*0,07*	7.159		253,4	0,000000
217	PLANS	*122*	*0,08*	11.199	0,01	246,9	0,000000
218	ISSUES	*123*	*0,08*	11.456	0,01	245,9	0,000000
219	ESTABLISHING	*64*	*0,04*	2.131		245,4	0,000000
220	DEVELOPED	*127*	*0,08*	12.388	0,01	244,1	0,000000
221	ANNUAL	*104*	*0,07*	7.897		244,1	0,000000
222	WOMEN	*207*	*0,13*	32.517	0,04	243,3	0,000000
223	DISSEMINATE	*31*	*0,02*	119		243,2	0,000000
224	AQUACULTURE	*23*	*0,01*	15		242,6	0,000000
225	SEAS	*48*	*0,03*	863		239,5	0,000000
226	CONVENTION	*74*	*0,05*	3.411		239,4	0,000000
227	UNDP	*26*	*0,02*	48		236,0	0,000000
228	REHABILITATION	*47*	*0,03*	835		235,6	0,000000
229	GROUPS	*151*	*0,10*	18.689	0,02	232,1	0,000000
230	ESTABLISHMENT	*77*	*0,05*	4.015		231,9	0,000000
231	STRENGTHENED	*50*	*0,03*	1.085		231,8	0,000000
232	COMMUNICABLE	*26*	*0,02*	54		231,1	0,000000
233	INCLUDE	*134*	*0,09*	14.839	0,02	229,7	0,000000
234	PARTICULAR	*181*	*0,12*	26.697	0,03	229,5	0,000000
235	NATURAL	*130*	*0,08*	14.044	0,02	228,1	0,000000
236	APPROACHES	*70*	*0,05*	3.214		227,0	0,000000
237	BODIES	*93*	*0,06*	6.748		225,6	0,000000
238	BASED	*178*	*0,12*	26.362	0,03	224,7	0,000000
239	IDENTIFY	*80*	*0,05*	4.744		222,5	0,000000
240	CONDITIONS	*132*	*0,09*	14.931	0,02	221,8	0,000000
241	DISEASES	*56*	*0,04*	1.777		219,7	0,000000
242	ASSESSMENTS	*48*	*0,03*	1.084		218,9	0,000000
243	REGULATORY	*50*	*0,03*	1.248		218,6	0,000000
244	INTERSECTORAL	*19*	*0,01*	5		217,9	0,000000
245	UNSUSTAINABLE	*26*	*0,02*	74		217,5	0,000000
246	RENEWABLE	*36*	*0,02*	376		216,5	0,000000
247	POLICY	*168*	*0,11*	24.532	0,03	215,6	0,000000
248	PROMOTION	*68*	*0,04*	3.261		215,1	0,000000
249	SYSTEMATIC	*54*	*0,04*	1.677		214,1	0,000000
250	ACCORDANCE	*57*	*0,04*	1.993		213,4	0,000000

N	WORD	FREQ.	AG21.LST %	FREQ.	RITT.LST %	KEYNESS	P
251	ASSESS	*62*	*0,04*	2.581		212,2	0,000000
252	ENHANCING	*36*	*0,02*	414		209,9	0,000000
253	RISKS	*61*	*0,04*	2.527		209,3	0,000000
254	SOCIAL	*223*	*0,14*	41.635	0,05	207,6	0,000000
255	PROVIDING	*89*	*0,06*	6.903		205,6	0,000000
256	PROJECTS	*80*	*0,05*	5.449		202,9	0,000000
257	SITU	*34*	*0,02*	365		202,7	0,000000
258	INCREASE	*130*	*0,08*	15.871	0,02	202,6	0,000000
259	DECISION	*137*	*0,09*	17.602	0,02	202,6	0,000000
260	BIOTECHNOLOGIE+	*17*	*0,01*	3		200,1	0,000000
261	TAKING	*147*	*0,10*	20.415	0,02	199,8	0,000000
262	GENETIC	*53*	*0,03*	1.829		199,7	0,000000
263	SAFE	*86*	*0,06*	6.725		197,5	0,000000
264	TECHNIQUES	*81*	*0,05*	5.845		197,3	0,000000
265	POVERTY	*63*	*0,04*	3.083		197,0	0,000000
266	KNOWLEDGE	*122*	*0,08*	14.415	0,02	196,5	0,000000
267	INDUSTRY	*139*	*0,09*	18.816	0,02	194,3	0,000000
268	AVERAGE	*99*	*0,06*	9.453	0,01	193,8	0,000000
269	ECOSYSTEM	*27*	*0,02*	153		192,9	0,000000
270	MINIMIZE	*31*	*0,02*	291		192,6	0,000000
271	COMBATING	*28*	*0,02*	185		192,2	0,000000
272	COPYPAGE	*15*		0		191,4	0,000000
273	YOUTH	*75*	*0,05*	5.135		189,5	0,000000
274	ROLE	*133*	*0,09*	17.815	0,02	188,1	0,000000
275	REDUCE	*84*	*0,05*	6.796		188,0	0,000000
276	ADOPT	*55*	*0,04*	2.387		184,0	0,000000
277	RISK	*108*	*0,07*	12.040	0,01	184,0	0,000000
278	FACILITATING	*29*	*0,02*	261		182,4	0,000000
279	FAO	*25*	*0,02*	137		180,2	0,000000
280	AGREEMENTS	*56*	*0,04*	2.614		179,9	0,000000
281	CLIMATE	*57*	*0,04*	2.752		179,6	0,000000
282	CHEMICAL	*68*	*0,04*	4.458		177,0	0,000000
283	FACILITIES	*83*	*0,05*	7.145		176,9	0,000000
284	PUBLIC	*196*	*0,13*	37.671	0,04	174,8	0,000000
285	DATABASES	*40*	*0,03*	1.008		174,2	0,000000
286	PRIORITY	*59*	*0,04*	3.187		174,0	0,000000
287	OBJECTIVE	*67*	*0,04*	4.421		173,7	0,000000
288	POPULATION	*108*	*0,07*	12.782	0,01	173,6	0,000000
289	QUALITY	*120*	*0,08*	15.882	0,02	171,9	0,000000
290	PRODUCTIVITY	*48*	*0,03*	1.877		169,8	0,000000
291	THROUGH	*303*	*0,20*	76.427	0,08	167,5	0,000000
292	COORDINATING	*27*	*0,02*	263		165,9	0,000000
293	PRODUCTS	*95*	*0,06*	10.350	0,01	165,4	0,000000
294	CONTROL	*160*	*0,10*	28.654	0,03	158,1	0,000000
295	TRADE	*128*	*0,08*	19.390	0,02	157,2	0,000000
296	LIVELIHOODS	*20*	*0,01*	81		155,0	0,000000
297	PARTICIPATORY	*22*	*0,01*	132		154,9	0,000000
298	AREA	*167*	*0,11*	31.423	0,03	153,7	0,000000
299	ADVERSE	*38*	*0,02*	1.140		153,1	0,000000
300	REFORESTATION	*17*	*0,01*	33		153,0	0,000000

N	WORD	FREQ.	AG21.LST %	FREQ.	RITT.LST %	KEYNESS	P
301	HYDROLOGIC	*14*		7		152,0	0,000000
302	EFFECTS	*93*	*0,06*	10.925	0,01	150,6	0,000000
303	SUCH	*369*	*0,24*	106.546	0,12	150,1	0,000000
304	SPECIALIZED	*33*	*0,02*	756		149,6	0,000000
305	SAFETY	*80*	*0,05*	8.064		149,4	0,000000
306	ASSIST	*48*	*0,03*	2.432		147,1	0,000000
307	TRANSPORTATION	*29*	*0,02*	507		146,2	0,000000
308	CONVENTIONS	*37*	*0,02*	1.173		145,2	0,000000
309	EQUITABLE	*30*	*0,02*	588		144,8	0,000000
310	II	*83*	*0,05*	9.055		144,3	0,000000
311	LANDS	*45*	*0,03*	2.109		144,2	0,000000
312	IMPROVEMENT	*58*	*0,04*	4.086		143,7	0,000000
313	AQUATIC	*27*	*0,02*	412		143,1	0,000000
314	TOTAL	*113*	*0,07*	16.705	0,02	142,9	0,000000
315	AGROFORESTRY	*15*		21		142,6	0,000000
316	MAKERS	*45*	*0,03*	2.183		141,4	0,000000
317	METHODS	*81*	*0,05*	8.833		140,9	0,000000
318	WATERSHED	*23*	*0,01*	227		140,7	0,000000
319	MOUNTAIN	*56*	*0,04*	3.885		140,2	0,000000
320	MOBILIZE	*20*	*0,01*	122		140,2	0,000000
321	INCREASING	*77*	*0,05*	8.040		139,4	0,000000
322	CENTRES	*63*	*0,04*	5.189		139,0	0,000000
323	TRANSNATIONAL	*26*	*0,02*	388		138,9	0,000000
324	COORDINATED	*25*	*0,02*	334		138,7	0,000000
325	CONSIDERATIONS	*45*	*0,03*	2.265		138,4	0,000000
326	EFFICIENCY	*54*	*0,04*	3.641		137,9	0,000000
327	ENSURING	*42*	*0,03*	1.895		137,5	0,000000
328	CLEANER	*33*	*0,02*	921		137,4	0,000000
329	MAKING	*152*	*0,10*	28.998	0,03	137,0	0,000000
330	III	*62*	*0,04*	5.130		136,4	0,000000
331	PRIVACY	*34*	*0,02*	1.040		135,7	0,000000
332	SUPPORTING	*49*	*0,03*	2.924		135,7	0,000000
333	REQUIREMENTS	*65*	*0,04*	5.762		135,3	0,000000
334	LINKAGES	*22*	*0,01*	215		135,0	0,000000
335	SUPPLY	*81*	*0,05*	9.315	0,01	133,9	0,000000
336	CONTRIBUTE	*46*	*0,03*	2.553		133,4	0,000000
337	OTHER	*439*	*0,29*	141.184	0,16	133,0	0,000000
338	TRENDS	*42*	*0,03*	2.034		132,1	0,000000
339	AFFORESTATION	*18*	*0,01*	94		131,3	0,000000
340	UNDERTAKEN	*46*	*0,03*	2.648		130,4	0,000000
341	CAPABILITY	*31*	*0,02*	847		130,3	0,000000
342	INTEGRATING	*25*	*0,02*	410		129,1	0,000000
343	UNSERVED	*12*		7		128,2	0,000000
344	DEPLETION	*24*	*0,02*	362		127,7	0,000000
345	EVALUATE	*33*	*0,02*	1.079		127,6	0,000000
346	IEEAS	*10*		0		127,6	0,000000
347	SEALEVEL	*10*		0		127,6	0,000000
348	PRECAUTIONARY	*19*	*0,01*	137		127,4	0,000000
349	PROVISION	*74*	*0,05*	8.207		126,7	0,000000
350	AMONG	*127*	*0,08*	22.612	0,02	126,5	0,000000

N	WORD	FREQ.	AG21.LST %	FREQ.	RITT.LST %	KEYNESS	P
351	INDICATORS	*33*	*0,02*	1.110		125,9	0,000000
352	SYSTEM	*188*	*0,12*	43.017	0,05	125,6	0,000000
353	AGRO	*15*		43		125,3	0,000000
354	IRRIGATION	*22*	*0,01*	281		124,0	0,000000
355	CONSIDER	*85*	*0,06*	11.064	0,01	123,9	0,000000
356	POTENTIAL	*84*	*0,05*	10.867	0,01	123,3	0,000000
357	MATERIALS	*66*	*0,04*	6.661		123,1	0,000000
358	INITIATE	*26*	*0,02*	539		122,8	0,000000
359	HARMONIZED	*14*		32		122,2	0,000000
360	PRIVATE	*108*	*0,07*	17.547	0,02	121,6	0,000000
361	SERVICES	*128*	*0,08*	23.579	0,03	121,5	0,000000
362	CURRICULA	*21*	*0,01*	254		120,5	0,000000
363	G	*93*	*0,06*	13.477	0,01	120,4	0,000000
364	ENHANCEMENT	*23*	*0,01*	378		118,7	0,000000
365	DRYLANDS	*11*		6		118,3	0,000000
366	CONSIDERATION	*57*	*0,04*	5.107		117,7	0,000000
367	FORMULATION	*29*	*0,02*	861		117,4	0,000000
368	SOIL	*52*	*0,03*	4.174		117,1	0,000000
369	FOOD	*108*	*0,07*	18.097	0,02	116,8	0,000000
370	ILLEGAL	*41*	*0,03*	2.345		116,7	0,000000
371	ACHIEVING	*37*	*0,02*	1.791		116,4	0,000000
372	INITIATIVES	*37*	*0,02*	1.832		114,9	0,000000
373	PATTERNS	*59*	*0,04*	5.669		114,9	0,000000
374	SEWAGE	*27*	*0,02*	721		114,6	0,000000
375	ESSENTIAL	*71*	*0,05*	8.516		112,6	0,000000
376	ENABLE	*53*	*0,03*	4.592		112,4	0,000000
377	FARMERS	*54*	*0,04*	4.790		112,4	0,000000
378	AVAILABLE	*132*	*0,09*	26.170	0,03	112,2	0,000000
379	INVENTORIES	*18*	*0,01*	169		111,8	0,000000
380	CHAPTER	*93*	*0,06*	14.595	0,02	109,4	0,000000
381	DISASTERS	*23*	*0,01*	474		108,9	0,000000
382	BASEL	*15*		81		108,5	0,000000
383	MUNICIPAL	*28*	*0,02*	914		108,4	0,000000
384	FRAMEWORKS	*18*	*0,01*	188		108,2	0,000000
385	FACTORS	*69*	*0,04*	8.400		107,8	0,000000
386	SCIENCES	*37*	*0,02*	2.045		107,6	0,000000
387	MULTIDISCIPLIN+	*18*	*0,01*	192		107,5	0,000000
388	CONSISTENT	*43*	*0,03*	3.010		107,0	0,000000
389	ACHIEVE	*60*	*0,04*	6.374		106,9	0,000000
390	PESTICIDES	*23*	*0,01*	500		106,5	0,000000
391	MINIMIZING	*15*		89		105,9	0,000000
392	EXPAND	*34*	*0,02*	1.676		105,9	0,000000
393	REDUCTION	*51*	*0,03*	4.556		105,5	0,000000
394	APPROACH	*95*	*0,06*	15.621	0,02	105,3	0,000000
395	INPUTS	*25*	*0,02*	706		103,5	0,000000
396	TERRESTRIAL	*20*	*0,01*	333		102,7	0,000000
397	EXPERTISE	*39*	*0,03*	2.520		102,5	0,000000
398	CORPORATIONS	*28*	*0,02*	1.029		102,2	0,000000
399	NEED	*186*	*0,12*	47.176	0,05	101,6	0,000000
400	DISSEMINATING	*14*		75		101,5	0,000000

N	WORD	FREQ.	AG21.LST %	FREQ.	RITT.LST %	KEYNESS	P
401	MULTISECTORAL	*9*		3		101,4	0,000000
402	INTERREGIONAL	*10*		9		101,4	0,000000
403	DEGRADED	*17*	*0,01*	184		101,1	0,000000
404	MAINTAIN	*53*	*0,03*	5.221		101,0	0,000000
405	INTERDISCIPLIN+	*19*	*0,01*	293		100,3	0,000000
406	APPLICATION	*71*	*0,05*	9.541	0,01	100,0	0,000000
407	SUPPORTIVE	*25*	*0,02*	767		99,7	0,000000
408	OPPORTUNITIES	*54*	*0,04*	5.551		99,1	0,000000
409	ONGOING	*23*	*0,01*	605		98,3	0,000000
410	SPECIAL	*109*	*0,07*	20.907	0,02	97,5	0,000000
411	ILO	*13*		61		97,3	0,000000
412	ASSESSING	*30*	*0,02*	1.386		96,9	0,000000
413	SUSTAINABLY	*10*		13		96,2	0,000000
414	FARMING	*36*	*0,02*	2.275		96,1	0,000000
415	PEST	*21*	*0,01*	472		96,0	0,000000
416	EDUCATIONAL	*54*	*0,04*	5.760		95,8	0,000000
417	LEGAL	*82*	*0,05*	12.964	0,01	95,6	0,000000
418	ATMOSPHERIC	*23*	*0,01*	646		95,5	0,000000
419	CONTACTS	*34*	*0,02*	2.005		94,9	0,000000
420	ASSEMBLY	*51*	*0,03*	5.175		94,7	0,000000
421	LIVESTOCK	*25*	*0,02*	854		94,7	0,000000
422	FORMULATE	*21*	*0,01*	488		94,6	0,000000
423	ASPECTS	*59*	*0,04*	7.006		94,5	0,000000
424	INDUSTRIALIZED	*19*	*0,01*	347		94,2	0,000000
425	ECOLOGICALLY	*15*		136		94,2	0,000000
426	ENTERPRISES	*32*	*0,02*	1.736		94,1	0,000000
427	ENDOGENOUS	*16*	*0,01*	182		93,6	0,000000
428	SPECIES	*69*	*0,04*	9.605	0,01	93,5	0,000000
429	ENFORCEMENT	*29*	*0,02*	1.352		93,2	0,000000
430	AVAILABILITY	*33*	*0,02*	1.914		93,1	0,000000
431	IV	*39*	*0,03*	2.898		93,0	0,000000
432	IPCS	*9*		7		93,0	0,000000
433	CHAP	*26*	*0,02*	996		92,9	0,000000
434	ADDRESSING	*26*	*0,02*	997		92,9	0,000000
435	HARMONIZE	*12*		50		92,4	0,000000
436	FOLLOWING	*123*	*0,08*	26.344	0,03	92,2	0,000000
437	MONITOR	*34*	*0,02*	2.100		92,1	0,000000
438	IDENTIFYING	*30*	*0,02*	1.519		91,9	0,000000
439	INTERNATIONALL+	*23*	*0,01*	706		91,7	0,000000
440	GOALS	*47*	*0,03*	4.533		91,2	0,000000
441	EARTHWATCH	*10*		18		91,2	0,000000
442	ARID	*19*	*0,01*	389		90,2	0,000000
443	COMMODITY	*25*	*0,02*	944		90,0	0,000000
444	OBSERVATION	*38*	*0,02*	2.868		89,6	0,000000
445	COLLECTION	*60*	*0,04*	7.666		89,2	0,000000
446	ALLEVIATION	*12*		59		88,8	0,000000
447	INVOLVEMENT	*44*	*0,03*	4.073		88,4	0,000000
448	PROCESS	*108*	*0,07*	21.935	0,02	88,4	0,000000
449	TRADITIONAL	*68*	*0,04*	9.831	0,01	88,2	0,000000
450	FRAGILE	*24*	*0,02*	870		88,2	0,000000

N	WORD	FREQ.	AG21.LST %	FREQ.	RITT.LST %	KEYNESS	P
451	AIMED	41	0,03	3.515		87,7	0,000000
452	COORDINATE	19	0,01	420		87,4	0,000000
453	EXTENSION	40	0,03	3.337		87,4	0,000000
454	GROUNDWATER	17	0,01	288		86,7	0,000000
455	CRITERIA	41	0,03	3.567		86,7	0,000000
456	OZONE	28	0,02	1.395		86,6	0,000000
457	PARTICIPATE	30	0,02	1.686		86,3	0,000000
458	REGULATIONS	43	0,03	4.005		86,0	0,000000
459	DISASTER	36	0,02	2.681		85,7	0,000000
460	HABITATS	19	0,01	451		84,9	0,000000
461	FORESTRY	24	0,02	939		84,9	0,000000
462	ENCOURAGING	36	0,02	2.732		84,5	0,000000
463	REGARD	50	0,03	5.669		83,8	0,000000
464	PREPAREDNESS	13		108		83,7	0,000000
465	ACTIONS	46	0,03	4.776		83,7	0,000000
466	RECOGNIZING	19	0,01	470		83,4	0,000000
467	ORGANS	24	0,02	1.013		81,5	0,000000
468	PREVENTIVE	18	0,01	419		81,1	0,000000
469	ADOPTING	22	0,01	799		80,8	0,000000
470	LIVELIHOOD	16	0,01	279		80,8	0,000000
471	COMPREHENSIVE	39	0,03	3.512		80,2	0,000000
472	REDUCING	36	0,02	2.939		80,0	0,000000
473	OPERATIONAL	28	0,02	1.609		79,4	0,000000
474	HAZARDS	21	0,01	736		78,5	0,000000
475	THEIR	614	0,40	247.711	0,27	78,2	0,000000
476	STRATOSPHERIC	12		97		77,9	0,000000
477	COMPLIANCE	26	0,02	1.376		77,6	0,000000
478	BIOMASS	14		188		77,6	0,000000
479	COMMENDATION+	32	0,02	2.334		77,3	0,000000
480	ADDRESS	51	0,03	6.401		77,3	0,000000
481	UNDERUTILIZED	7		3		77,1	0,000000
482	COOPERATIVE	19	0,01	564		76,9	0,000000
483	BASIC	67	0,04	10.793	0,01	76,3	0,000000
484	FORUMS	12		105		76,1	0,000000
485	PRONE	22	0,01	897		76,1	0,000000
486	ADMINISTRATIVE	38	0,02	3.533		76,1	0,000000
487	VULNERABLE	32	0,02	2.388		76,1	0,000000
488	ARRANGEMENTS	47	0,03	5.543		75,8	0,000000
489	UNDERTAKING	27	0,02	1.588		75,5	0,000000
490	RADIOACTIVE	21	0,01	800		75,3	0,000000
491	PRINCIPLES	47	0,03	5.595		75,1	0,000000
492	MILLION	106	0,07	23.477	0,03	75,0	0,000000
493	SHELTER	26	0,02	1.468		74,6	0,000000
494	ENHANCED	29	0,02	1.938		74,5	0,000000
495	MITIGATE	13		161		74,0	0,000000
496	INDUSTRIAL	69	0,04	11.696	0,01	73,5	0,000000
497	H	58	0,04	8.560		73,5	0,000000
498	ULTRAVIOLET	14		222		73,2	0,000000
499	SCIENTISTS	37	0,02	3.513		72,8	0,000000
500	FUNDING	39	0,03	3.986		71,9	0,000000